COMMON AGENCY THEORY AND THE INDUSTRIAL ORGANIZATION OF HEALTH CARE

By

WILLIAM EDWARD ENCINOSA, III

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1995

ACKNOWLEDGMENTS

This work was supported financially by a Jim Walter-Lanzillotti Dissertation Fellowship. I thank the Public Policy Research Center at the University of Florida for their generous support.

I am indebted to David Sappington for his insightful comments, invaluable advice, and constant support in the difficult task of translating economic theory into public policy. I am also grateful to Tracy Lewis, Jon Hamilton, Mike Ryngaert, and Steve Slutsky for many helpful discussions. In particular, I thank David Martimort of IDEI, France, for his help with common agency theory. I also thank Professor Louis Block of the Mathematics Department for providing me with very generous computer facilities.

This work has benefited greatly from the comments of seminar participants at Harvard University's John F. Kennedy School of Government, Carnegie Mellon's Heinz School of Public Policy, and IUPUI. I am also grateful for the many helpful discussions with the faculty and scholars participating in the 1995 Robert Wood Johnson Health Policy Research Programs at the University of California-Berkeley, the University of California-San Francisco, and the University of Michigan.

Finally, I am especially grateful to my parents for their continuous support and encouragement during this work.

TABLE OF CONTENTS

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

COMMON AGENCY THEORY AND THE INDUSTRIAL ORGANIZATION OF HEALTH CARE

By

William Edward Encinosa, III

May 1995

Chairman: Professor David E. M. Sappington
Major Department: Economics

Common agency theory deals with a competitive version of principal-agent theory in which multiple principals contract with the same agent. This dissertation develops common agency theory for applications to the industrial organization of the health care industry.

In the first essay, the efficiency of three health care systems is examined. In a multi-payer system the public payer (Medicare) uses a mix of prospective payments and pass-through payments, while the private payer (a managed care insurer) uses a quality-based reimbursement rate through utilization review. Cost-shifting in the multi-payer system induces the hospital to overinvest in technology, even under private insurer utilization reviews. Furthermore, pass-through payments of capital are scaled back in equilibrium since they create a moral hazard problem that allows the hospital to goldplate, i.e, to invest in wasteful, non-technological capital. This may explain Medicare's current policy to phase-out pass-through payments. In a single-payer system investment in technology is curtailed, and goldplating can arise. When trilateral negotiations between the payers and the hospital are admitted, technological efficiency results when a uniform quality-based reimbursement rate is negotiated. Moreover, this all-payer system is immune to goldplating.

The second essay studies the interrelationship between a hospital's capital structure and the payment plans designed by Medicare and the private insurance sector to reimburse the hospital's cost of capital and technology. To counter the opportunism of a managed care private insurance sector engaging in utilization review, the hospital will use debt when bankruptcy is costly.

The final essay derives the equilibrium hospital contract with a physician group that has private information on the risk rating of the health maintenance organization it serves. Bundled physician charges are then derived when hospitals compete for an exclusive contract.

CHAPTER 1
GENERAL INTRODUCTION

The United States is presently undergoing a national health care crisis. In 1994, expenditures on health care exceeded \$ 1 trillion, more than 14 percent of GDP. Medical costs are projected to soar to 19 percent of GDP by the year 2000 (Clinton 1993). Per capita, the United States spent 40 percent more on health care than Canada, the second highest spender, and twice as much as the major European countries in 1991 (Starr 1994). By 1990, the United States was spending more on health care than on education and defense combined. Yet, paradoxically, about 18 percent of Americans do not have medical insurance. In fact, 26 percent of Americans had no health insurance coverage at some time between 1987 and 1989 (Starr 1994). Moreover, 86 percent of these uninsured were members of working households (Roberts 1993). Among OECD countries, the United States ranks nineteenth in infant mortality and twenty-first in life expectancy for men (Clinton 1993). Thus, for its extremely high level of medical expenditures, the United States does not seem to be delivering a commensurate level of health care to its people.

This crisis has lead to an unprecedented public policy debate over federal reform of the health care industry. The last major federal health reform to occur was in 1965 with the establishing of Medicare to insure the elderly and the severely disabled. At that time, health care was only 6 percent of GNP. Yet, now Medicare spending is running 23 percent higher than the rate of inflation (Clinton 1993). Consequently, the Medicare Trust Fund will be empty by the year 2002. The Reagan Administration did make substantial changes in Medicare payments to hospitals with the introduction of the Prospective Payment System (PPS) in 1984. Under this system, hospitals receive a fixed payment for each patient in a specific diagnostic group, instead of a cost-based reimbursement for each service provided to the patient. This

held Medicare reimbursement expenditures down initially. However, after 1989, hospital expenditures began to soar again, accounting for about 40 percent of all medical expenditures in the United States.

Parallel to this reform in the public sector, the private insurance sector and the hospital industry are undergoing major transformations. Over the past 5 years, more than 400 hospitals have sought mergers. More importantly, there has been a significant movement to merge insurance and health care provision. These mergers have taken the form of health care maintenance organizations (HMOs) and preferred provider organizations (PPOs). Over half of the urban population is projected to be enrolled in HMOs by 1998. A key feature of these managed care facilities is that they engage in utilization review. That is, they monitor the appropriateness of medical treatment and review patient outcomes. This prevents doctors and hospitals from performing unnecessary, expensive procedures. Arnold Relman, former editor of *The New England Journal of Medicine*, estimates that roughly one third of health care expenditures are medically unnecessary. It appears that HMOs have experienced an initial one time savings of about 10 percent. Beyond that, managed care cost advantages seem to be waning. Thus, we see a failure in both the private and public sectors of the health care market to contain medical costs.

The following three essays in this dissertation will take an industrial organizational approach to examining medical cost containment incentives under both regulation and competition. Industrial organization provides a set of tools to analyze these changes in markets and the consequences of public policy initiatives. In particular, we will focus on the the tool of *common agency theory.* This is the theory of multi-principal agent theory, which investigates how two rival principals should design incentive contracts for an agent that they happen to share. This scenario arises quite frequently in the health care industry. For example, the first two essays consider Medicare as one principal and the private insurance sector as the rival principal. Both compete in reimbursing the same hospital (the agent) for treating their patients. In

contrast, the third essay takes two rival hospitals to be the principals competing for the exclusive patronage of an HMO (the agent). Using this common agency approach, I am able to gain new important insights into the industrial organization of the health care industry.

The first essay, titled *Cost-Shifting and Goldplating in Health Care*, examines the efficiency of three health care systems. First, I show that a single-payer system under PPS (similar to Canada's health system), in which the government is the sole medical insurer, induces the hospital to underinvest in technology and to undersupply treatment quality. However, I show that this single-payer system can be made efficient by implementing a mixed payment in which a cost-based element is added to the PPS rate.

Next, I examine a multi-payer system in which part of the market is insured by Medicare (under PPS), while the remainder of the market is insured by a Managed Care private insurance sector (e.g., HMOs). This system induces the hospital to overinvest in technology and to undersupply treatment quality. This reflects the current state of the United States' multi-payer system. By using an inverse elasticity rule in its reimbursing of technology and operating costs, Medicare is able to shift operating costs to the private insurance sector. Even though the private insurer is allowed to engage in aggressive utilization reviews of the hospital, this is not enough to counter the distortion caused by Medicare's cost-shifting. Thus, overinvestment in technology persists under utilization review. This result dispels a common notion in health care that managed care will mitigate Medicare's cost-shifting and bring the multi-payer system back to efficiency.

Another commonly held belief is the idea that a mixed payment system (PPS with a cost-based element) will correct the distortions in the multi-payer system as it did in the single-payer system. In fact, I show the exact opposite occurs. Mixed payments will actually exacerbate the both the technology and quality distortions in

the multi-payer system. In essence, the mixed payment just gives Medicare another instrument to shift costs to the private insurance sector.

Moreover, I show that the multi-payer system induces the hospital to goldplate, i.e., to invest in wasteful capital such as fancy artwork and lobby waterfalls. To deter goldplating, Medicare must scale back its pass-through payment of capital. This may partially explain Medicare's 1991 decision to phase-out pass-through payments of capital.

Finally, I introduce the all-payer system in which the private and public insurance sectors trilaterally negotiate with the hospital over a uniform reimbursement rate. By instituting a uniform rate structure, this all-payer system prevents Medicare from cost-shifting. Moreover, by dismantling the pass-through payment, this system is immunized against goldplating.

The second essay, *Optimal Hospital Capital Structure*, introduces a stochastic version of the model in the first essay. This enables me to examine how for-profit hospitals should finance their investments under the risk arising from the highly volatile nature of hospital admissions. I show that the hospital will issue some debt in order to counter the private insurer's opportunistic behavior in utilization reviews. By using debt, the hospital purposely exposes itself to a positive risk of bankruptcy. Since bankruptcy proceedings are costly, this forces the private insurer to increase its reimbursement rate in order to reduce the risk of bankruptcy. This rate increase for the hospital outweighs the risk of using debt. Thus, I predict that even as for-profit hospitals lose their tax-shelter advantage with debt, hospitals will continue to issue bonds in order to mitigate the ever-increasing practice of utilization review by the managed care sector.

In the last essay, titled *Pricing under Exclusive Dealing*, I examine how hospital competition for an exclusive contract with a physician group affects the physician group's pricing of services to an HMO. This common agency model is much more advanced than the first two essays in that I now allow the agent to have private

information. Specifically, the physician group (the agent) has private information on the HMO's demand for health care. The hospitals (the principals) do not know the HMO's demand, and so must design incentive contracts to induce the physicians to reveal this information. Each hospital offers a reimbursement rate to the physician group for the use of the hospital's technology. The physician group than bundles this hospital rate with a physician reimbursement rate and charges this price bundle to the HMO.

I examine how hospital competition and integration affects the physician-insurer bargaining over prices and utilization levels. I derive the equilibrium hospital contracts for two rival hospitals competing through the same shared physician group which is privately informed about the insurer's bargaining power. For substitute (complementary) medical procedures, competition through a common physician group results in higher (lower) physician prices and less (more) severe under-utilization of medical services when compared with a multi-specialty hospital merger. Under strong substitiutes, the hospitals instead compete head-to-head for the exclusive services of the physician group. For an insurer with a weak (strong) bargaining position, exclusive dealing results in lower (higher) physician prices and under-utilization (over-utilization) when compared with a hospital vertically integrated with the physician group.

CHAPTER 2
COST SHIFTING AND GOLDPLATING IN HEALTH CARE

2.1 Introduction

Health care reform has been one of the most hotly debated public policy issues during the Clinton Administration. It is widely agreed that any health care reform must first address cost containment. In 1994, expenditures on health care in the U.S. are expected to exceed $ 1 trillion, more than 14 percent of GDP. Medical costs are projected to soar to 32 percent of GDP by 2020. Many health care economists (Evans (1986), Newhouse (1988), and Weisbrod (1991), for example) argue that the present system of financing high-tech medical technology is responsible for the rapid explosion in medical costs. According to the former long-time editor of the *New England Journal of Medicine*, Arnold Relman, technology is the "engine behind the rise in medical costs," driven by the excessive number of doctors "trained to provide high-tech, expensive services"(Relman (1989)). In the debate over how to deal with this overinvestment in technology, controversy often erupts over the merits of three different health care systems: the single-payer system, the multi-payer system, and the all-payer system. In this paper we offer an economic analysis of these three systems of health care and provide answers to the following key questions: (1) Which of the three systems best provides the hospital with the incentive to invest in the socially efficient level of technology? (2) What is the role of price negotiations in these systems? (3) How do the incentives of cost-based reimbursement, prospective reimbursement, mixed reimbursement, and quality-based reimbursement differ among the three systems? (4) To what extent should the government make pass-through payments of hospital capital in each system? and (5) Which system is susceptible to hospital goldplating?

The United States' health care system is currently a multi-payer system, in which the public and private insurers set reimbursement rates non-cooperatively. A common criticism of multi-payer systems is that the public insurer (Medicare, Medicaid) is able to shift costs to the private insurance sector. While such cost shifting has long been criticized for being unfair or inequitable for patients, it is only recently that cost-shifting has been shown to lead to economic inefficiencies such as decreased patient hospital length of stay (Glazer and McGuire (1994)) and possible technological underinvestment (Ma and McGuire (1993)). However, it has yet to be demonstrated how cost-shifting may induce excessive investment in high-tech medical equipment. The goal of this paper is to link technological overinvestment to cost-shifting. This explanation of overinvestment is in contrast to the traditional "medical arms race" (MAR) explanation in which it is argued that hospitals compete by providing too many high-tech medical services. The implication is that capital is wasted, leading to higher costs without commensurate benefits (Robinson and Luft (1985), Kopit and McCann (1988), and McManis (1990)). This seems to be the prevailing explanation for hospital overinvestment. In fact, antitrust judges have even embraced this idea to the extent of permitting hospital mergers on the grounds that it would bring the medical arms race to an end [1]. Over the last five years, more than 400 hospitals sought mergers[2]. Nevertheless, hospital costs continue to escalate, suggesting that incentives for overinvestment may not be due entirely to competition among hospitals[3]. This paper reveals and analyzes a more fundamental source of the overinvestment problem. We show that excessive hospital investment can result from the cost-shifting induced by competition among payers, even in the absence of any competition among hospitals.

[1]For example, in a recent decision to permit a merger between the two largest hospitals in Roanoke, Virginia, the district court judge wrote, "As a general rule, the hospital rates are lower, the fewer the number of hospitals in an area" (*United States v. Carilion Health System* 892 F2d 1042), cited in Dranove and White (1994).

[2]*The Tampa Tribune*, July 1, 1994.

[3]Moreover, Dranove et al. (1992) and Zwanziger and Melnick (1993) do not find significant empirical evidence of a MAR.

Some argue that the government's cost-shifting in a multi-payer system may be counteracted by the rapid upsurge of "Managed Care" plans in the private sector (such as health maintenance organizations and preferred provider organizations). However, a key finding here is that hospital overinvestment in technology persists even in such a managed care environment, even if private insurers can set aggressive quality-based reimbursement rates through utilization review. That is, after a utilization review of the hospital's quality and technology, the private insurer can update its reimbursement rate. This reflects the current "Outcomes Movement," which Relman (1988) dubbed the "third revolution in medical care." The basic goal of the outcomes movement is to link reimbursement rates data on patient outcomes and hospital quality. An interesting result of our paper is that these quality-based, or outcomes-based, hospital rates cannot resolve the overinvestment problem of the multi-payer system. That is, these aggressive payments do not provide strong enough incentives to counter the cost-shifting problem. The reason for this conclusion is that with a first-mover advantage, the government can impose a cost-shifting externality on the bargaining process between the hospital and the private payer by using a combination of two payments: (1) a prospective payment, which is a fixed payment per patient, and (2) a pass-through payment which pays for a percentage of capital and technology. To compensate for this externality, the hospital must always overinvest in technology in order to improve its bargaining position with the private insurance sector.

Interestingly, Medicare used such a combination of pass-through payments and prospective payments up until 1991, when it decided to phase-out pass-through payments over a ten year period. Why would Medicare phase-out pass-through payments, since doing so would seem to diminish its ability to shift costs to the other payer? To answer this, we show that the pass-through payment is susceptible to hospital goldplating—investment in unproductive, non-technological capital, such as elaborate artwork for private rooms, lobby waterfalls, plush offices and athletic facilities

for doctors, etc. While the investment in high-tech medical equipment is excessive, it is still a productive asset in that it has value to patients. Goldplating, on the other hand, is completely wasteful in that it has no value to patients. To deter goldplating, Medicare must scale back its pass-through payments, trading-off diminished cost-shifting for less goldplating. However, overinvestment in technology as well as some goldplating may still persist in the multi-payer equilibrium.

How might these persistent problems with the multi-payer system be solved? A popular idea is to replace the multi-payer system with a single-payer system, in which the government is the only medical insurer[4]. While such a change would eliminate cost-shifting, it would also, however, severely curtail socially beneficial investments in technology. A large strand of the literature has recently suggested that the underivestment problem of prospective payments can be prevented by mixing the prospective payment with a cost-based retrospective payment. In contrast, we show that this restoration of efficiency will not occur in reality since the cost-based element of the mixed payment is susceptible to goldplating. As a result, goldplating as well as underinvestment in technology persist in the single-payer equilibrium.

To solve the problems of the multi-payer and single-payer systems,we introduce an all-payer system. Here, the public payer, the hospital, and the private payer trilaterally negotiate a uniform reimbursement rate which can vary with delivered quality. All-payer systems are used in Japan, Europe, and in four American states. However, these do not typically use quality-based reimbursement. This is unique to our all-payer model. The key result of our paper is that this particularly aggressive all-payer system is technologically efficient. Moreover, this all-payer system is immune to cost-shifting and hospital goldplating.

This chapter is organized as follows. In Section 2, the three health care systems are introduced. In Section 3, the equilibrium reimbursement plans and investment

[4]In 1993, approximately one third of the House Democrats signed on to a bill that would establish a single-payer system.

levels under the three health care systems are derived. Mixed payments are analyzed in Section 4. Goldplating is then introduced in Section 5. Finally, in Section 6, the public policy implications of the analysis for medical cost containment are discussed. All proofs are relegated to the Appendix.

2.2 The Model

First, we set up a model of a multi-payer health care system. To begin, we consider the demand side of the health care market. For simplicity, we assume that consumers in the health care market are fully insured[5]. Thus, market demand is assumed to be price inelastic with respect to the price of medical services. However, following Rogerson (1994) and Ma (1994), we assume that the hospital's demand for admissions X(I) is directly influenced by the quality of care or intensity of care I offered by the hospital. ProPAC's (1993) working definition of 'intensity' is the number and complexity of patient care resources, or intermediate outputs, used in producing a patient care service. Intensity I may include length of stay, services per admission, special amenities, etc.

Unlike Ma and McGuire (1993), we do not assume that that hospital volume is a function of the hospital's level of technology T. Recent research (e.g., Dranove and White (1992)) suggests that technology usually does not attract patients to the hospital directly, but instead attracts physicians to the hospital. This results in a physician-driven 'medical arms race' (MAR) in which hospitals overinvest in high-tech equipment just to attract the best doctors. This phenomenon is the traditional explanation for excessive investment in hospital technology. However, recent empirical evidence indicates that the MAR effect has diminished or disappeared altogether (see Zwanziger and Melnick (1993)). Thus, it remains to determine what else besides

[5]We assume that patients obtain insurance from only one insurer at a time. For a discussion of the problems in supplying medical insurance, see Diamond (1992), Newhouse (1994), and Lewis and Sappington (1994).

a MAR may be driving the overinvestment in medical technology. A central finding of this paper is that payer competition can cause overinvestment in technology, even in the absence of a MAR, and even under prospective payments.

In order to investigate payer competition in isolation, we first remove the incentive for a MAR by assuming that the hospital's number of physicians with admitting privileges is fixed exogenously. Thus, the hospital will have no demand-side motivation for over-capitalization through technology (however, demand will be induced by hospital intensity I). Instead, the hospital's only incentive for overinvesting in technology will be shown to arise from the supply-side effects of payer competition.

On the supply side, the hospital has an operating cost of c(I,T) per patient. There are no fixed costs associated with supplying I. The hospital's only fixed cost is T. Following Ma and McGuire (1993), n will be the fraction of patients that are insured by the public payer (Medicare, Medicaid); the other (1-n) of the market is insured by the private payer[6]. The public payer pays a reimbursement rate α per patient to cover c(I,T) and to help pay for T. In addition, the public payer can make a pass-through payment pT that is independent of the number of discharges, where p$\in [0,1]$. This payment plan represents the Prospective Payment System (PPS) administered currently by Medicare and by some state Medicaid programs (ProPAC 1993). The private payer will pay a rate β per discharge. β is determined by negotiations between the private payer and the hospital. Historically, pass-through payments have never been instituted by private payers, and so will not be considered here.

Our model of the health care system focuses on the interactions among three actors: the public payer, the hospital, and the private payer[7] . The public payer and the private payer will compete in setting their rates (α, p) and β, respectively, in order to reimburse the hospital's operating cost c(I,T) and cost of technology T. However,

[6]We are essentially modeling the multi-payer system as a von Stackelberg common agency, where the insurers are the principals and the hospital is the agent. Our model departs from the usual common agency model by allowing the hospital to negotiate with one of the principals.

[7]See Encinosa (1994) for a model in which there is a fourth player: the capital market.

this payer competition will further entail opportunistic behavior on the part of each payer.

First, we describe how the public payer might be able to shift some of the costs of treating Medicare patients on to the private payer. A well known problem in health care is that public insurers, such as Medicare and Medicaid, often do not reimburse the hospital for the fair share of their patients' costs, thereby shifting a burden of unpaid costs to the private insurance sector[8]. To model the public payer's ability to shift costs in this manner, we give the public payer the first-mover advantage as a price setter. Thus, in stage 1 of the four-stage game, the public payer (Medicare) sets the margin α and the pass-through payment parameter p. Medicare is presumed unable to update its payment plan after the hospital selects I and T, perhaps due to a lack of managerial resources[9]. Anticipating the results of the next two stages of the game, the public payer will select a reimbursement plan (α, p) in order to maximize the net social benefit to the segment of the market it serves:

$$M(\alpha, p) = n[U(I,T) - \alpha X(I)] - pT,$$

where U(I,T) is the gross social benefit when the hospital provides technology level T and treatment intensity level I. Note that an incomplete contracts approach is adopted. We assume that the insurers cannot write complete *ex ante* contracts contingent on the hospital's technology and treatment intensity. State-of-the-art innovation in medical technology and medical procedures often advance so rapidly that it is usually difficult, if not impossible, for the private insurance sector to characterize precisely all relevant aspects of a hospital's technology in a contract.

In stage 2, anticipating the per patient fee that the private payer will negotiate (β) in stage 3, the hospital chooses the level of investment T and intensity I in order

[8]The American Hospital Association estimates that Medicaid payments covered only about 80% of average costs in 1991. Hospital profit margins on Medicare's PPS patients averaged -4% in 1991 (AHA 1992).

[9]At times, the budgets for U.S. Veteran Health Administration hospitals have been updated according to hospital volume. However, this is not the case for Medicare and Medicaid payments in general (Newhouse (1994)).

to maximize profits:[10]

$$Y(I,T) = [n\alpha + (1-n)\beta]X(I) - c(I,T) + pT - T.$$

While public payer cost-shifting has been a major focus of recent research, opportunism on the part of private payers has been neglected in the literature. The salient feature of this paper is that it introduces two forms of private payer opportunism that have recently emerged with the rapid upsurge of "Managed Care" plans (such as preferred provider organizations (PPOs) and health maintenance organizations (HMOs))[11]. A key feature of these managed care plans is that they often engage in utilization review. That is, they monitor the hospital's choice of quality I and level of technology T. This gives the private payer the ability and flexibility to quickly update rates in response to the hospital's behavior. Thus, the aggressive monitoring of utilization review allows the private payer to behave opportunistically by basing its rate β on the observed I and T. In order to avoid a complex dynamic game between the hospital and the private payer, we assume that the hospital's choice of I and T in Stage 2 is immutable. One possible justification for the hospital to commit to T and I is that it may be very costly for the hospital to constantly adjust T and I in a dynamic pricing game. For example, to induce the full level of demand X(I) may require a period of commitment to the level I. That is, the doctors must build a relationship with their patients over a period of time in which they are credibly committed to a guaranteed level of care I. In addition, it may be costly for the hospital to quickly purchase or resell high-tech equipment on the spot market.

Like the prospective payment α, the payment β is a fixed rate per patient. However, unlike α, the payment β may depend explicitly on I and T. Thus, to distinguish β from a prospective payment, we will call β a *quality-based reimbursement rate.*

[10]We assume that the hospital cannot discriminate by payer in their intensity provision. See Rogerson (1994) for such a model in the single payer case.

[11]Over half of the urban population is projected to be enrolled in HMOs by 1998.

We assume the private payer has bargaining power vis a vis the hospital. Such power may arise in practice when, for example, the private insurer forms large networks of doctors and patients across several markets. To capture the private payer's purchasing power, β is assumed to be determined in stage 3 as the solution to a generalized Nash bargaining game[12]. Presuming the private insurance sector to be competitive, the private insurer's objective is to maximize the net expected social benefit to the patients it represents:

$$V(\beta) = (1-n)[U(I,T) - \beta X(I)].$$

Note that if the hospital leaves the bargaining table, it derives income only from serving the patients insured by the government. That is, the hospital's threat point is endogenously determined by the public insurer's reimbursement plan. However, we assume that the investment T is not hospital-specific and is completely reversible and redeployable. This is characteristic of hospital technology. In such a case, Encinosa (1994) shows that private owners of the hospital will always redeploy the hospital's capital to other markets if the hospital makes the out-of-equilibrium decision to not negotiate with the private insurance sector. Consequently, the hospital's threat point will always be zero. Next, the private insurer has no purchasing power with any outside hospital. Since the private payer cannot obtain a discount elsewhere, we will normalize the private insurer's threat point to 0. With a (0,0) threat point, the generalized Nash bargaining solution β^* maximizes $Y^q V^{1-q}$ while guaranteeing $V,\ Y\ \geq 0$, where $q \in [0,1]$ is the relative bargaining power of the hospital.

In the final stage 4, demand X(I) is realized (no rationing is allowed). The public insurer pays the hospital $\alpha X(I) + pT$, while the private insurer reimburses the hospital $\beta X(I)$.

When $n \in (0,1)$, a *multi-payer system* emerges in which the public and private payers set their reimbursement rates non-cooperatively. When $n = 1$ we have the

[12]See Ellis and McGuire (1990) for an interesting Roth-Nash bargaining model in which the quantity of services is negotiated rather than price.

single-payer system, in which the government is the sole insurer and sets hospital prices. We shall refer to the $n = 0$ case as the *privatized system*. We will not focus on the privatized system since it is a polar case of the much more general all-payer system. The *all-payer system* embodies a multi-payer system in which the government, the hospital, and the private insurer engage in trilateral negotiations to set a uniform quality-based rate without a pass-through payment.

More specifically, instead of the government setting its rates before everyone else in stage 1, the government, together with the hospital and private insurer, negotiates a rate ϕ after the hospital selects T and I. That is, in the all-payer system, the trilateral Nash bargaining solution ϕ^* solves the program

$$\max_\phi Y^q(I,T)V^x(\phi)M^z(\phi,0) \text{ such that } Y,V,M \geq 0,$$

where $n,q,x,z \in [0,1]$ such that $q+x+z=1$. Note that when n=0 the all-payer system is equivalent to the privatized system. This all-payer model is very similar to the all-payer systems found in Japan and in four American states in which a uniform fee is negotiated by all payers. However, what is unique about our all-payer model is that we consider a quality-based uniform rate. That is, the fundamental differences between the single-payer system and the all-payer system in our model are that the all-payer system sets the uniform price after observing both T and I, and, moreover, prohibits pass-through payments. We assume that it is beyond the ability of the government to monitor I and T in the the single-payer system. We take this view of the single-payer system since it is most reflective of the Canadian single-payer system in which reimbursement rates are purely prospective and are rarely based on I and T (Merrill (1994), p. 251).

2.3 The Overinvestment Problem

For each of the three health care systems, we will derive the subgame-perfect equilibrium investment level of the four-stage game and compare it to the socially

optimal investment level. The socially efficient investment level T^E and intensity level I^E maximize the total expected social welfare $U(I,T) - c(I,T)X(I) - T$, which we assume to be concave in I and T. Before comparing the systems, we will define

$$\Delta(I,T|n) \equiv (1-n)U(I,T) - c(I,T)X(I) + n\alpha X(I) + (p-1)T$$

to be the total surplus that is to be bargained over by the hospital and the private insurer in the multi-payer system. Δ is assumed to be concave in I and T. Next, it will be helpful to construct the following elasticities. The *bargaining elasticity of technology* T is defined as

$$\epsilon_{TT} = \epsilon_{TT}(I,T|\alpha,p) = \frac{-T\Delta_{TT}}{\Delta_T}. \tag{2.1}$$

Similarly, the *bargaining cross-elasticity of intensity* I is defined as

$$\epsilon_{IT} = \epsilon_{IT}(I,T|\alpha,p) = \frac{I\Delta_{TI}}{\Delta_T}. \tag{2.2}$$

To see that ϵ_{TT} is an elasticity, observe that equation (1) can be rewritten as

$$\epsilon_{TT} = \left(\frac{-d(\Delta_T)}{\Delta_T}\right) \Big/ \left(\frac{dT}{T}\right).$$

Thus, ϵ_{TT} gives the percent change in Δ_T, the marginal bargaining value of technology T, over the percent change in T. Similarly, ϵ_{IT} is the percent change in the marginal bargaining value of technology, when the treatment intensity I is changed, over the percent change in I. Throughout the paper we will maintain the following comparison of elasticities[13].

<u>*Assumption 1*</u> $\epsilon_{IT} > \epsilon_{TT}$.

Assumption 1 will hold whenever treatment intensity I and technology T are sufficiently strong operating cost complements. For most inpatient procedures it is quite reasonable that I and T are cost complements ($c_{IT} < 0$). That is, the marginal

[13]In equilibrium ($\Delta_T = 0$), Assumption 1 should be interpreted as $\frac{\epsilon_{IT}}{\epsilon_{TT}} > 1$.

operating cost of an additional unit of treatment intensity decreases as the medical equipment accompanying the procedure becomes more advanced. If I and T are operating cost substitutes, then Assumption 1 will hold if I and T display sufficient utility complements ($U_{IT} > 0$). Next, we need the following additional assumption.

Assumption 2 $\frac{U}{U_I} \geq \frac{X}{X_I}$.

This assumption is sufficient to ensure that the $\Delta \geq 0$ constraint is non-binding in equilibrium. This will simplify the equilibrium analysis in the following comparison of the three health care systems.

Proposition 1 *When pass-through payments are allowed:*

1. *In the single-payer system equilibrium* (T_S, I_S, p_S), *the hospital will underinvest in technology* $(T_S < T^E)$ *and will provide less than the socially efficient level of treatment intensity* $(I_S < I^E)$. *Moreover, the government will choose not to use pass-through payments* $(p_S = 0)$. *The hospital will earn nonnegative profits of* $q\Delta(I_S, T_S|1)$.

2. *Under bargaining, the all-payer system will provide the hospital with the incentive to invest efficiently at* T^E *and to supply the efficient intensity* I^E, *regardless of its negotiating power* q. *The hospital will earn nonnegative profits of* $q\Delta(I^E, T^E|0)$.

3. *In the multi-payer system equilibrium,* (T_0, I_0, p_0), *there exists a critical market share* $n_0 \in (0,1)$ *such that the hospital will overinvest in technology* $(T_0(n) > T^E)$ *if* $n < n_0$. *The hospital will undersupply treatment intensity* $(I_0(n) < I^E)$, *regardless of its negotiating power* q. *When* $n < n_0$, *the government will use a pass-through payment to pay for* p_0 *percent of the hospital's technology, where*

$$p_0 = \min\left\{ 1 \ , \ T_0\Delta_{TT} + \frac{X}{X_I}\Delta_{TI} + nU_T \right\}. \tag{2.3}$$

The hospital will earn nonnegative profits of $q\Delta(I_0, T_0|n)$.

These results are illustrated in Figure 1. Since the MAR-effect has been removed by decoupling demand from technology, the hospital has no incentive to invest in technology (beyond some mandatory minimum threshold) under a single-payer system. The reason for this is that the payment system is "too prospective". To induce investment in technology, a complete pass-through (p=1) is not enough, the government would also have to reimburse a portion of operating costs retrospectively (this will be addressed in the next section). Similarly, the payment is also still too prospective with respect to the demand-inducing treatment intensity. This is a common result in the literature (Glazer and McGuire (1994), Ellis and McGuire (1986,1990)).

However, Proposition 1 reveals a more interesting consequence of prospective payments that warrants special attention. Among the three systems, overinvestment in technology is only induced by prospective payments under the multi-payer system. The overinvestment results from externalities created by the uncoordinated dual regulation of the hospital. To see this, first note that the hospital now has an incentive to invest in technology in order to increase its bargaining position with the private insurer. Next, since the investment is not hospital-specific, ex post negotiations do not result in the hold-up problem of Klein, Crawford, and Alchian (1977) and Williamson (1975,1977). Thus, the private insurer cannot expropriate the investment. Therefore, under the all-payer system, this investment is socially efficient since the public payer cannot externally influence the bargaining process.

However, under the multi-payer system, the public payer imposes an externality on the negotiation process. More precisely, the public payer engages in cost-shifting by employing an "inverse elasticity rule." That is, since technology is more inelastic than intensity ($\epsilon_{IT} > \epsilon_{TT}$), the public payer increases the pass-through p (with respect to the single-payer p_S) in order to increase reimbursement of T and lowers α to decrease the reimbursement for I. As a result, the hospital now has the incentives to overinvest in technology and to undersupply treatment intensity.

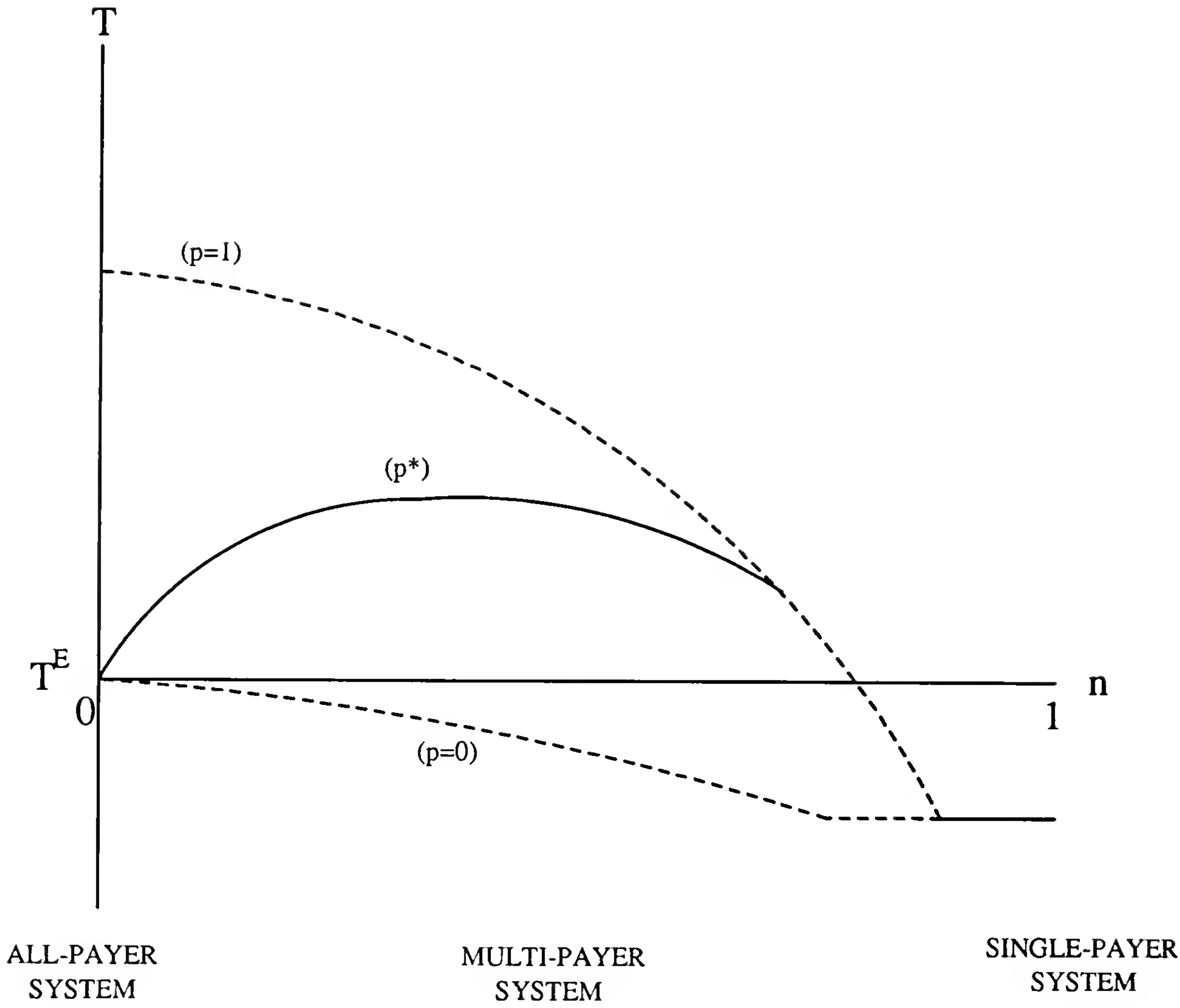

FIGURE 2-1 : EQUILIBRIUM INVESTMENT

This result may well characterize the present U.S. health care market under prospective rate setting. It is widely agreed that prospective payments from Medicare as well as aggressive payments from HMOs induce too low of a treatment intensity, while technology costs continue to soar (Newhouse (1992), Thorpe (1992)). As potential evidence of this overinvestment problem in multi-payer systems and the underinvestment problem in single-payer systems, it is interesting to note that Atlanta has more Magnetic Resonance Imaging facilities (MRIs) than all of Cananda (*Newsweek*, July 25, 1994). At the beginning of Medicare's Prospective Payment System (PPS) in 1984, Atlanta had only one MRI facility. Now it has at least thirty (Eckholm and Pear (1993)). Overall, the U.S. has eight times more MRIs than does Canada on a per capita basis (Starr (1994)). Finally, if we interpret I as patient length of stay, it is interesting to note that the U.S. has the lowest average length of stay among OECD nations, with an average of 7.1 days, compared to 12 to 24 for most other nations (Reinhardt (1992)).

Finally, we note how the results of Proposition 1 stand in contrast to the results of Ma and McGuire (MM) (1993). MM assume that intensity is exogenous, that operating costs are constant, that demand is induced by technology (an MAR-effect), and that the private payer can set rates only before the hospital selects T. Under these assumptions, MM find that a multi-payer system results in technological underinvestment. Moreover, MM's single-payer system is efficient. These two results do not seem to support the general empirical evidence. In contrast, our characterizations of the multi-payer and single-payer systems seem to be more realistic.

Moreover, in MM's model, the government will implement complete pass-through payments (p=1) in both the multi-payer and single-payer system. In our single-payer model the government completely dismantles pass-through payments. In addition, our multi-payer model supports the use of an interior pass-through $p \in (0, 1)$ in equilibrium. In practice, Medicare used a .85n pass-through parameter prior to 1991. The public policy implications of our equilibrium characterization of the pass-through

payment will also diverge from Ma and McGuire's policy recommendations. This will be explained more fully in Section 6.

2.4 Mixed Payment Systems

In the preceding section all government reimbursements of operating costs were paid prospectively. That is, a fixed per patient rate was set in advance. In Proposition 1, we saw that this led to equilibrium rates that were "too prospective" in the single-payer system, leading to underinvestment, even when the cost-based pass-through payments could be made. In this section we show that the single-payer system can be efficient when the government is allowed to use mixed payments in which a portion of operating costs can be reimbursed retrospectively. In the spirit of Ellis and McGuire (1986), *mixed payment plans* are defined by the expanded payment plans (α, p, r) and (β, R), in which the public insurer pays an additional rC(I,T) per patient and the private insurer pays an additional RC(I,T) per patient. As the next Proposition demonstrates, these mixed reimbursement rates have very different consequences under each health care system.

Proposition 2 *Under mixed payments*[14]*:*

1. *The single-payer system induces the hospital to invest efficiently in technology and to provide the socially efficient level of treatment intensity. The government's payment is completely cost-based:* $\alpha = 0$, $p = 1$, *and* $r = 1$.

2. *In the multi-payer system equilibrium under mixed payments,* (T_1, I_1, p_1), *there exists* $n_1 \in (0,1)$ *such that the hospital will select* T^E *and* I^E *for* $n > n_1$, *but will overinvest at the level* $T_1(n) > T_0(n) > T^E$ *and will under-supply intensity,* $I_1(n) = I_0(T_1)(n) < I^E$ *when* $n < n_1$. *Moreover,* $r > 0$ *and* $R = 0$.

[14] In parts (2) and (3) of Proposition 2, the private payer is actually indifferent about using R. We assume the mild refinement that the payers will select the R=0 equilibrium. In Proposition 4, this refinement will no longer be necessary due to goldplating.

3. *The all-payer system admits an equilibrium that is efficient and uses no cost-based payments* ($R = 0$).

This Proposition reveals an important disparity between the single-payer system and the multi-payer system when the government employs mixed payments. Mixed payments resolve the underinvestment problem that arises when prospective payments alone are employed in the single-payer system (recall Proposition 1). Not only is it in the government's best interest to use partially cost-based payments, but it is socially optimal. In contrast, the government's actions are no longer aligned with society's best interest under a multi-payer system. Instead, the government will now over-employ r, the cost-based element of the payment, in order to shift costs to the private insurer. In fact, r gives the public payer even more cost-shifting leverage than it had under the prospective payment plan. As a result, the overinvestment problem of Proposition 1 is exacerbated under mixed payments.

The main result of this section is that mixed payments increase the degree of cost-shifting in the multi-payer system. Glazer and McGuire (GM) (1994) find a similar result in a related multi-payer model. In the GM model, fixed costs (technology) are exogenous so that pass-through payments are unnecessary. Furthermore, intensity does not induce demand. Moreover, actual costs are not contractible; only allocated costs are contractible. In this setting, GM show that both payers use a mixed payment (in which the cost-based rate is approximated by a cost-allocation formula), inducing the profit-maximizing hospital to under-supply intensity. Here, when technology is endogenous and actual costs are contractible, the hospital will overinvest in technology and still under-provide treatment intensity. This finding reveals that cost-shifting does not necessarily result from an inability to contract on actual costs. Cost-shifting can arise when actual costs are contractible if the private payer does not use a mixed payment in response to the government's mixed payment. In practice, private insurers rarely use mixed payments even though some of the rival public-sector contracts are mixed.

Our result on the efficiency of mixed payments in the single-payer system is not new. Mixed payments have been advocated quite regularly for single-payer systems (Ellis and McGuire (1986,1993), Goodall (1990), Ma (1994), Pope (1989)). However, in the next section we show that this traditional result may be overturned when goldplating is possible.

2.5 Goldplating

In general, T may often include not just expenditures on technology, but the entire capital costs of constructing a new facility to house the technology (e.g., a new wing on the hospital for an open-heart surgery center). So far we have considered T to be a productive asset in that it had patient value U(I,T). In the preceding analysis, it was implicitly assumed that the government monitors the hospital's investment T to insure that non-productive hospital assets with no value to patients are not bought and reimbursed through the direct pass-through payment. However, in reality Medicare and state Medicaid agencies often do not have the resources to monitor the capital expenditures of hospitals. Indeed, state certificate-of-need laws, mandated by the National Health Planning and Resources Development Act of 1974, required states to establish agencies to regulate hospital investments (Simpson (1985)). However, many states have since abandoned certificate-of-need laws since they have failed to prevent excessive investment (Sloan (1988), Salkever and Bice (1976), Joskow (1981))[15].

This problem is not restricted to the state level. " 'We've set up a system that in no way rewards prudent behavior,' says Gail Wilensky, former chief of the Health Care Financing Administration (HCFA), which runs Medicare. 'In fact, the more you spend, the more you get.' And it doesn't much matter what the money buys.

[15] With the expiration of Pub. L. No. 93-641, twelve states abandoned certificate-of-need laws altogether (Hackey (1993)).

(Medicare) pays as willingly for a lobby waterfall as for an inner-city emergency room"(*Business Week*, April 22, 1991).

When a hospital's capital expenditures go unmonitored while being largely reimbursed with direct pass-through payments, the hospital may find it advantageous to invest in capital that has value to management and doctors but not to patients. For example, hospitals often build elaborate offices for doctors adjacent to the hospital. Other examples of goldplating include elaborate architectural facades, overly plush doctors' lounges, on-site athletic centers and child-care centers for the hospital staff, elaborate artwork for private rooms, and general managerial slack. Unfortunately, the patients do not receive any direct benefits from these managerial resources, even though they may be subsidized by Medicare.

One way to resolve this moral hazard problem is to set up proper managerial incentives and ex post utilization reviews (as analyzed in Encinosa and Sappington (1995), for example). However, the next Proposition shows that goldplating can be deterred simply by restricting pass-through payments, without adversely distorting the technological investment incentives. Let us assume that a non-productive hospital capital expenditure of G has value $\psi(G)$ to management and zero value to patients. Moreover, we will suppose that a hospital which is vertically integrated with the payers would never invest in G. That is, we assume G has negative net present value in isolation, i.e., $\psi'(0) \leq 1$ and $\psi'' < 0$[16]. Note that if $\psi \equiv 0$, the hospital will choose G=0. That is, the hospital will never *waste* resources. However, when $\psi > 0$, then the hospital will choose a positive G under a multi-payer system. This is referred to as *abusing* resources, following Blackmon (1992)[17].

<u>Proposition 3</u> *If capital expenditures are not monitored by the public payer and if mixed payments are prohibited, then:*

[16]We are assuming that operating costs are independent of G. An interesting problem that we do not consider is the further inability of the government to easily distinguish operating costs from capital costs (see Merrill (1994)).

[17]Moreover, since $\psi' < 1$, abuse does not include non-technological expenditures that the hospital would make if it paid its full costs, such as lobbying expenditures and political contributions.

1. *In the multi-payer system,* $p^* = 1 - \psi'(G^*)$ *if there exists a positive* G^* *that solves the equation*

$$\psi''(G)[1 - p_0 - G\Delta_{TT}] - \psi''(G)\psi'(G) - \Delta_{TT}\psi'(G) + \Delta_{TT} = 0, \qquad (2.4)$$

where p_0 *is the pass-through payment in the monitored case of equation (3). Moreover, the hospital will underinvest in technology if pass-through payments are prohibited.*

2. *There is no goldplating (G=0) in the single-payer and the all-payer systems.*

In general, it is quite difficult to verify whether equation (4) has a positive solution. However, if ψ is quadratic, then we have the following existence result.

<u>Corollary 1</u> *If* $\psi(G) = G - \frac{b}{2}G^2$ $(b > 0)$ *in the multi-payer system (without mixed payments), then the hospital will abuse resources at the level* $G^* = \frac{p_0}{b - 2\Delta_{TT}} > 0$. *Moreover, the government will scale back its pass-through payment to* $p^* = bG^* < p_0$, *forcing the hospital to scale back technology from the* T_0 *level.*

Goldplating arises only in the multi-payer system since pass-through payments are instituted only in the multi-payer system when mixed payments are prohibited (see Proposition 1). Note that in isolation, the return on asset G is always negative (since $\psi' - 1 < 0$). However, once a positive pass-through payment is introduced, there is always an asset level G that earns a positive return ($\psi'(G) - 1 + p > 0$). If the public payer cannot monitor the hospital's investment in order to weed out unproductive assets, Corollary 1 indicates that it is then optimal for the government to scale back its pass-through payment. Note that the pass-through payment is never fully dismantled. Thus, goldplating, or abuse, is never completely eradicated in equilibrium. The reason for this is that the government's ability to shift costs is eroded without the use of the pass-through payment. In essence, the moral hazard problem forces the public payer to trade-off some cost-shifting for a partial deterrence of goldplating. Not only is goldplating partially deterred, but there is an additional

merit to scaling back pass-through payments: the hospital will also scale back its overinvesting in technology.

In Proposition 3, goldplating did not arise in the single-payer system since pass-through payments are not used in that system when mixed payments are prohibited (see Proposition 1). However, recall from Proposition 2 that the government will implement a complete pass-through ($p = 1$) of all capital in the single-payer system once a mixed payment is allowed. As a consequence,the single-payer system is susceptible to goldplating under mixed payments.

<u>*Corollary 2*</u> *Under mixed payments, we have $G > 0$ and $p, r \in (0,1)$ in the single-payer system, resulting in underinvestment in technology and under-provision of treatment intensity.*

The results of Corollary 2 stand in contrast to the recent strand of literature which suggests that single-payer systems are efficient under mixed payments (Ellis and McGuire (1986, 1993), Goodall (1990), Ma (1994), Pope (1989)). These studies do not address goldplating. The salient finding of Corollary 2 is that if goldplating is a problem as the empirical evidence indicates, then mixed payments cannot restore efficiency in the single-payer system. To deter goldplating, the government must scale back both its cost-based rate and its pass-through payment. The government must trade-off less technology and less treatment intensity for less goldplating. Together, Proposition 1 and Corollary 2 reveal that a single-payer system will persistently curtail socially efficient technological investment and treatment intensity.

2.6 Public Policy Implications

When the Prospective Payment System was first introduced in 1983, it is conceivable that Medicare failed to anticipate the incentive for hospital goldplating that would arise due to the pass-through payment. The HCFA's attempts at regulatory reform of Medicare's pass-through payment in 1986 and 1987 were blocked by

Congress due to intense hospital lobbying. It was not until 1991 that HCFA succeeded in persuading Congress to force Medicare to phase-out pass-through payments over a ten year period. Corollary 1 provides a possible theoretical justification for this HCFA policy of prohibiting the cost-based pass-through payment. Although many have thought that the HCFA's primary motivation for a phase-out of pass-through payments was to end a federally financed "medical arms race", we have shown that the reform actually deters goldplating as well as eases the overinvestment externality of dual regulation that persists in the absence of an "arms race".

Note that Corollary 1 reveals that it is not in HCFA's best interest to lobby for a full dismantlement of the pass-through payment. While it is indeed optimal for HCFA to scale back the pass-through payment, it is always optimal for HCFA to pass-through a small fraction of capital costs. Indeed, it is even socially beneficial to maintain some degree of pass-through payments. While a complete prohibition on pass-through payments will completely deter all goldplating, it will nevertheless induce the hospital to underinvest in technology and under-provide treatment intensity.

Note that this policy recommendation does not go as far as Ma and McGuire (1993), who advocate a complete restoration of the pass-through payment ($p=1$) in order to ease the severity of the underinvestment. Goldplating is not consider in MM's model. If it were, their model might also recommend an interior pass-through p [18].

While the pass-through payment reduction of Corollary 1 does not necessarily lead to efficiency in technology and intensity in the multi-payer and single-payer systems, it is interesting to note that the hospital would perform at the efficient levels if the public payer charged the hospital a fixed licensing fee (or entry fee). Moreover, the public payer would benefit.

[18] However, even at an interior p, underinvestment in technology would still likely result, in contrast to the overinvestment in technology of Corollary 1.

<u>Proposition 4</u> *If the public payer is allowed to impose a fixed licensing fee L on the hospital for any purchase of technology, then:*

1. *The single-payer system induces the hospital to invest at the socially efficient technology level T^E and to provide the efficient level of intensity I^E under an equilibrium fixed license fee $L = Y(I^E, T^E)$.*

2. *The multi-payer system induces the hospital to invest at the socially efficient technology level T^E and to provide the efficient level of intensity I^E under an equilibrium fixed license fee $L = \Delta(I^E, T^E|n)$. Moreover, the public payer will still use a positive pass-through payment with L.*

This Proposition is not surprising. As expected, the public payer uses L to extract all of the surplus from the hospital and the private insurance sector. This full rent extraction leads to efficiency since L allows (α, p) to be adjusted to essentially sell-out Medicare to the hospital and the private insurance sector. This sell-out equivalently resets n to 0. Thus, as in the all-payer case (or privatized case), it is in the interest of the hospital to perform efficiently (Proposition 1). However, Medicare benefits because it can then extract all of the surplus from the sell-out via the license L. Note that efficiency is obtained without the private payer charging a licensing or entry fee.

Interestingly, this licensing fee policy is in contrast to the two-part hospital tariff policies of Gal-Or (1994) and Ma and Burgess (1993). Gal-Or shows that overinvestment may be the result of hospital competition under stochastic demand in a single-payer setting. Ma and Burgess demonstrate that suboptimal intensity may result from hospital competition. However, both papers reveal that efficiency can be restored if the single payer provides the hospitals with lump sum subsidies. In contrast, our model advocates the use of licensing fees when there is no hospital competition.

It is interesting to note that many states which have maintained certificate-of-need laws have required hospitals to pay entry fees or application fees for new capital

purchases. However, Proposition 4 may reveal why these certificate-of-need laws have failed to contain costs. First, the efficiency results of Proposition 4 work because it is the public payer who collects the entry fee. This allows the public payer to adjust (α, p) in a way that does impose an externality on the hospital. In contrast, many states have set up independent regulatory health systems agencies (HSAs) to control the licensing of new capital for hospitals. The HSA licensing fees are set independently of Medicare and Medicaid's reimbursement decisions. According to Reinhardt (1992), the HSAs are completely divorced from the reimbursement decision. Proposition 4 indicates that efficiency obtains only if Medicare has complete jurisdiction over setting entry fees and licensing fees.

A second problem with the HSAs is that they have tended to set fees that are substantially below the optimal fee L recommended by Proposition 4. For example, in Florida the application fee to request permission to open an open-heart surgery center is \$ 22,000. However, the hospital can make this back in the first 15 minutes of the first operation they do (*The Tampa Tribune*, September 24, 1994). The fixed fees required by Proposition 4 are extremely large in that they must extract the entire surplus from the hospital and the private insurance sector.

Such large licensing fees are obviously not feasible in reality. Instead, Proposition 1 offers a more viable alternative in order to obtain efficiency: the all-payer system. It is interesting to note that in 1977 the Carter Administration proposed a series of all-payer revenue controls on hospitals. However, after three years of legislative battles, the initiative was defeated by intense hospital lobbying (Ginsburg (1988)). About the same time, all-payer systems were tried successfully in Maryland, Massachusetts, New Jersey, and New York. Medicare, Medicaid, and private insurers cooperated in setting a uniform rate (Thorpe (1993)). Hospital costs were controlled somewhat until the 1980's when Medicare withdrew from these all-payer state systems in order to set its own rates (Merrill (1994), p.47). Medicare prefers a multi-payer system

since it can shift costs to the other payers. As we have shown, this cost-shifting under the multi-payer system will encourage excessive technological investment.

Although the uniform rate structures in the single-payer system and all-payer system remove this cost-shifting externality of the multi-payer system, only the all-payer system induces the hospital to perform efficiently. The reason for this is that the all-payer uniform rate is quality-based, linking the hospital's choice of I and T to its reimbursement. In fact, quality-based reimbursement has recently become an issue in the public sector. For example, Oregon recently passed legislation to link the approval of capital projects to the hospital's patient outcomes (Alter and Holtzman (1992)). In addition, New York's Commissioner of Health , Mark Chassin, recently proposed linking hospital reimbursement to quality-of-care measures in the 1993 renegotiations of the state's prospective hospital reimbursement system (Darby (1993)). The key result of this paper is these new aggressive quality-based reimbursement mechanisms will only work efficiently in the all-payer system in which cost-shifting and goldplating are both eradicated.

2.7 Conclusion

The results of our theoretical paper suggests some important directions for future empirical work. First, how has Medicare's phase-out of the pass-through payment affected hospital investment? Second, what is the degree of cost complements between I and T ($c_{IT} < 0$) and what is the degree of ϵ_{IT}? Lastly, how hospital-specific and redeployable is hospital technology? If hospital capital is irreversible, then the all-payer system will result in underinvestment due to Williamson's (1985) hold-up problem. In such a case, the multi-payer system may be more efficient, with the overinvesting problem mitigating the hold-up problem.

CHAPTER 3
OPTIMAL HOSPITAL CAPITAL STRUCTURE

3.1 Introduction

The most hotly debated public policy issue during the Clinton Administration is the financing of health care. It is widely agreed that any health care reform must first address cost containment. In 1990, expenditures on health care in the U.S. exceeded 12 percent of GNP. Medical costs are projected to soar to 36 percent of GNP by 2020. The key to successful reform is to understand the source of this rise in health care cost. Unfortunately, in attempting to explain this explosion in medical costs, many economists have failed to understand that the financing of health care is subject to a fundamental paradox. The health insurance market (composed of a public sector: Medicare and Medicaid, and a private insurance sector) attempts to set reimbursement rates to cover the hospital's cost of capital and technology. However, any reimbursement plan affects the hospital's expected future earnings, which, in turn, affects the hospital's cost of capital. The circularity of this process suggests that both the hospital and the health insurance market should take into account the simultaneous determination of the cost of capital and the prices of health care services. This paper lays a foundation for the equilibrium determination of hospital capital costs and medical insurance reimbursement.

This chapter is organized as follows. In Section 2, we explain how the capital market prices the hospital's financial securities. In Section 3, we examine the strategic role of hospital debt. All proofs are relegated to the Appendix.

3.2 The Model

To begin, we consider a stochastic version of Ma and McGuire's (1993) model of the demand side of the health care market. For simplicity, we assume that consumers in the health care market are fully insured[1]. Thus, market demand is assumed to be price inelastic with respect to the price of medical services. However, market demand depends crucially on the level of advanced technology and state-of-the-art care offered by the hospital. Letting T be the hospital's monetary level of investment in technology, we define X(T,z) to be the number of patients who demand admission to the hospital in the realized state of nature z. The demand for most types of acute medical services is highly stochastic. This volatility will be modeled by the demand shock $z \in [0,1]$, generated by the continuous distribution F(z). Higher levels of z correspond to greater demand so that X(T,z) is increasing in z for all T. We assume that it is illegal for the hospital to ration services. Thus, capacity (e.g., number of beds) is fixed at a sufficiently large enough level so that health care services are not rationed in equilibrium, even when demand is most pronounced (i.e., when z=1)[2]. Since there is no rationing, X(T,z) can also be considered as the realized number of discharges from the hospital. Demand for hospital services X(T,z) is assumed to be increasing and concave in technology investment T. Finally, n will be the fraction of patients that are insured by the public payer (Medicare, Medicaid); the other (1-n) of the market is insured by the private payer[3]. For simplicity, we do not allow patients to be covered by both insurers.

On the supply side, the hospital has a constant operating cost of c per discharge. The public payer pays c per discharge plus a margin α per discharge to help pay for

[1] We wish to show that overinvestment can occur under universal coverage. For a discussion of the problems in supplying medical insurance, see Diamond (1992) and Lewis and Sappington (1994).

[2] We are primarily concerned with the optimal T when there is a threat of bankruptcy when z is low. The threat of rationing when z is high is left for future research. Gal-Or (1994) investigates the optimal capacity for a hospital duopoly when investment T is fixed.

[3] While we do not allow rationing in general, we do allow the hospital to exit either the privately insured market or the entire market altogether.

T. In addition, the public payer can make a pass-through payment $p(T+i)$ that is independent of the number of discharges, where p$\in[0,1]$ and where i is the risk-free interest rate. This payment plan represents the Prospective Payment System (PPS) administered currently by Medicare and by some state Medicaid programs (ProPAC 1993)[4]. The private payer will pay c plus a margin β per discharge. β will be determined by negotiations between the private payer and the hospital. Historically, pass-through payments have never been instituted by private payers, and so will not be considered here.

Our model of the health care system focuses on the interactions among four actors: the public payer, the hospital, the capital market, and the private payer. The timing of the game among these players is as follows. In stage 1, the public payer (Medicare) institutes the margin α and the pass-through payment parameter p. In stage 2, the hospital chooses the level of investment T, and a mix of equity and debt to externally finance the investment T by issuing new shares and bonds. At stage 3, the capital market determines the market value of the hospital's securities. Finally, in stage 4, the private payer's margin β arises out of negotiations between the hospital and the private payer as the Nash bargaining solution. Then the demand shock z is realized, medical services are delivered as demanded, and payments are made, with bankruptcy being declared if necessary. The exact details of each stage will be discussed next.

Stage 1: The Government's Payment Plan.

The public payer is the first institution that sets prices, establishing the margin α and the pass-through payment parameter p. Medicare is presumed unable to update its payment plan after the hospital makes its investment, perhaps due to a lack of

[4]Prior to 1983, Medicare used a cost-based plan, paying rc for some rate r. See Wedig, et. al. (1988) and Sloan, et. al. (1988) for estimations of the cost of capital for hospitals under the cost-based plan.

managerial resources. Thus, while anticipating the results of the next three stages of the game, the public payer selects a reimbursement plan (α, p) in order to maximize the expected net social benefit to the segment of the market it serves:

$$M = M(\alpha, p) = n[U(T) - (c + \alpha)X(T)] - pT(1 + i), \text{ where } X(T) = \int_0^1 X(T, z)dF(z)$$

and where U(T) is the gross social benefit when the hospital provides technology level T. Note that i is the risk-free interest rate.

Stage 2: The Hospital's Financial Strategy

The hospital is assumed to maximize its profits. A nonprofit analog of our model can easily be constructed. We choose to focus on the for-profit case since the number of for-profit hospital beds increased by 41 percent between 1976 and 1986, while the total number of hospital beds declined by 10 percent during the same period (American Hospital Association (1987)).

The hospital is assumed to have no retained earnings and no previous debt obligations. It is owned by the shareholders of a previous issue of voting equity (common stock). These shareholders will be referred to as the "old" stockholders. Next, the hospital invests T dollars in technology and finances its investment from external sources. When new shares are issued to new shareholders, $s < 1$ will represent the fraction of the hospital's total number of shares that these new shares comprise. It is also possible for the hospital to repurchase some of its existing equity (i.e., $s < 0$). The hospital can also issue debt with a face value of D. The hospital chooses both the magnitude of the technology outlay $T = T(\alpha, p)$ and the capital structure $(s, D) = (s(\alpha, p), D(\alpha, p))$ to finance T. In doing so, the hospital anticipates the outcome of the next two stages of the game. The presumed autonomy for the hospital in choosing its investment and capital structure arises because, in practice, medical

insurers typically do not have the expertise to specify technology and security design for the hospital in detail.

The operating income of the hospital is

$$R(\alpha, \beta, p, T, z) \equiv [n\alpha + (1-n)\beta]X(T, z) + pT(1+i).$$

As the hospital takes on debt, there may be a critical value of the demand shock z^*, below which demand is too meager to enable the hospital to pay all of its debt. This critical value is defined by

$$z^* \equiv \min\{z \geq 0 \mid R(\alpha, \beta, p, T, z) \geq D\}. \tag{3.1}$$

When the demand for hospital services is sufficiently low that $z < z^*$, limited liability applies, the hospital declares bankruptcy, and the bondholders become the residual claimant. When demand is sufficiently high that $z \geq z^*$, the hospital remains solvent, and both the old and new shareholders remain the residual claimants. Note that $F(z^*)$ is the probability of bankruptcy.

Bankruptcy imposes costs on the bondholder due to legal fees, reorganization, and disruption of services. Moreover, a financially distressed hospital may cut back on the quality of care, increasing the risk of malpractice. Following Brander and Lewis (1988), we assume that bankruptcy costs are proportional to the size of the shortfall in the hospital's earnings from its debt obligation. That is, realized bankruptcy costs are $b[D(\alpha, p) - R(\alpha, \beta, p, T, z)]$ for some positive exogenous cost b per unit of shortfall when $z < z^*$. Since bondholders are also protected by limited liability, we assume that the hospital is liquidated whenever $b[D(\alpha, p) - R(\alpha, \beta, p, T, z)] > R(\alpha, \beta, p, T, z)$. Let z^{**} be the critical value below which the hospital is liquidated. When $z^* \geq z \geq z^{**}$, the hospital is not liquidated, but is reorganized under the ownership of the bondholders who receive $R(\alpha, \beta, p, T, z) - b[D(\alpha, p) - R(\alpha, \beta, p, T, z)]$. Thus, expected bankruptcy costs are

$$L(\alpha, \beta, p, T) \equiv \int_0^{z^{**}} R(\alpha, \beta, p, T, z)dF(z) \; + \; b\int_{z^{**}}^{z^*} [D(\alpha, p) - R(\alpha, \beta, p, T, z)]dF(z).$$

Given the hospital's risky debt obligation D, the total expected return on the financial securities is

$$\Pi(\alpha,\beta,p,T) = \int_0^1 R(\alpha,\beta,p,T,z)dF(z) - L(\alpha,\beta,p,T).$$

These expected profits are the combined expected returns to shareholders (old and new) and bondholders, divided between them according to their respective claims. However, hospital management will select T, s, and D in order to maximize the expected profit to the old shareholders [5]:

$$Y = Y(T,s,D,E,B) = (1-s)\int_{z^*}^1 [R(\alpha,\beta,p,T,z) - D(\alpha,p)]dF(z). \tag{3.2}$$

Stage 3: The Capital Market.

Next, the capital market establishes the market value E of the hospital's new equity s, and the market value B of its debt with face value D. We assume that the capital market is risk-neutral, competitive, and that investors correctly anticipate the outcome of the private insurer's negotiations with the hospital over the reimbursement margin β. The hospital's securities are priced fairly so that both the new shareholders and the bondholders earn an expected return equal to i, the risk-free interest rate. That is, in equilibrium, the capital market sets the following valuations:

$$\begin{aligned}(1+i)E &= s\int_{z^*}^1 [R(\alpha,\beta,p,T,z) - D(\alpha,p)]dF(z), \text{ and} && (3.3)\\ (1+i)B &= D(\alpha,p)(1-F(z^*)) + \int_0^{z^*} R(\alpha,\beta,p,T,z)dF(z) - L(\alpha,\beta,p,T) && (3.4)\\ \text{with } E+B &= T. && (3.5)\end{aligned}$$

While the market only requires $E + B \geq T$, we assume the private insurer can prohibit the hospital from accumulating discretionary cash flows. Thus, equation (5)

[5]We assume that there is no moral hazard problem between the management and the old shareholders. For the design of managerial investment incentives to resolve such a moral hazard problem, see Encinosa and Sappington (1993).

is assumed to hold. In equation (3), the value of the equity is equal to the fraction s of the expected operating income of the hospital that goes to the new stockholders, net of debt payments when the hospital is solvent. The first term on the right side of equation (4) is the expected return to bondholders when demand is high enough to ensure that they are paid D in full. The last two terms represent the expected operating income of the hospital when it becomes financially distressed due to low demand, net of bankruptcy costs. Combining equations (3),(4), and (5), we obtain the capital market equilibrium condition:

$$(1+i)T(\alpha,p) = D(\alpha,p)(1-F(z^*)) + \int_0^{z^*} R(\alpha,\beta,p,T,z)dF(z) - L(\alpha,\beta,p,T) + \\ +s\int_{z^*}^1 [R(\alpha,\beta,p,T,z) - D(\alpha,p)]dF(z) \quad (3.6)$$

The hospital's optimal capital structure (s, D) will always satisfy the capital market equation (6). Solving for s in equation (6), the hospital profits in equation (2) can now be expressed as

$$Y = \int_0^1 R(\alpha,\beta,p,T,z)dF(z) - T(1+i) - L(\alpha,\beta,p,T).$$

Stage 4: Negotiations with the Private Insurance Sector.

Next, we model the private insurer as a managed care facility, such as an HMO, engaging in utilization review. That is, we allow the private insurer to negotiate the margin β with the hospital *after* observing the hospital's investment T, capital structure (s, D), and the market valuation (E, B). To capture this aspect of negotiation, we will derive β as the generalized Nash bargaining solution. First, we specify each party's utility function. We assume that the private insurance sector is competitive so that the private insurer's objective is to maximize the net expected social benefit

to the patients it represents:[6]

$$V(\beta) = (1-n)[U(T) - (c+\beta)X(T)].$$

When the hospital management goes to the bargaining table, it will still represent only the old stockholders, seeking to maximize Y from equation (2)[7]. If the hospital backs away from the bargaining table, its only income will be from serving the patients insured by the government. That is, the hospital's threat point is endogenously determined by the public insurer's reimbursement plan.

Note that if the hospital fails to negotiate (i.e., $\beta = 0$), it will prefer to serve the publicly insured patients rather than to close the hospital. However, if the hospital serves only the public patients, then the capital market will devalue the hospital's securities from (E, B) to (E_0, B_0) (where the 0 subscript refers to the $\beta = 0$ case):

$$(1+i)E_0 = s\int_{z_0^*}^{1} (n\alpha X(T,z) + pT(1+i) - D)dF(z),$$

$$(1+i)B_0 = \int_{z_0^*}^{1} dF(z) + \int_0^{z_0^*} (n\alpha X(T,z) + pT(1+i))dF(z) - L_0, \text{ where}$$

$$L_0 = \int_0^{z_0^{**}} (n\alpha X(T,z) + pT(1+i))dF(z) + b\int_{z_0^{**}}^{z_0^*} [D - (n\alpha X(T,z) + pT(1+i))]dF(z),$$

$$z_0^* = \min\{z \geq 0 \mid n\alpha X(T,z) + pT(1+i) \geq D\}, \text{ and}$$

$$z_0^{**} = \min\{z \geq 0 \mid b[D - (n\alpha X(T,z) + pT(1+i))] \geq n\alpha X(T,z) + pT(1+i)\}.$$

Although the hospital will prefer to serve the public sector when $\beta = 0$, the bondholders and the new shareholders may instead want the hospital to liquidate T and pay back as much of B and E as possible. We will assume that the discounted resale value of T is γT, where $\gamma \in \{0,1\}$. When $\gamma = 1$ the technology is completely

[6]For simplicity, in V and M, we assume that patient welfare does not decline when bankruptcy occurs. However, welfare is affected indirectly by the risk of bankruptcy via β.

[7]In the stage 4 subgame, there is no conflict of interests between the old stockholders, the new stockholders, and the bondholders. They all desire β as large as possible. With more complex market structures, conflicts of interest may arise, as in Brander and Lewis (1986,1988). However, if the technology is hospital-specific, including the new investors in the bargaining process may force T to be scaled back in stage 2. This will be discussed later.

redeployable. In contrast, when $\gamma = 0$, the capital asset T is hospital-specific. In most cases medical technology is not hospital specific. However, the accompanying infrastructure (i.e., a new wing of the hospital) may be irreversibly sunk. Moreover, γ may include the litigation cost of suing the hospital to liquidate T. In particular, since $E_0 < E$ and $B_0 < B$, the bondholders and new shareholders will always prefer suing the hospital to liquidate T when $\gamma = 1$ (recall that $T = E + B$ from equation (5)). However, the investors cannot completely recover T when the asset T has no resale value or when litigation is too costly ($\gamma = 0$). In this case the investors will keep their downgraded securities E_0 and B_0.

Consequently, the old shareholders' threat point for the bargaining game is

$$r(\gamma) = (1-\gamma)(1-s)\int_{z_0^*}^{1} (n\alpha X(T,z) + pT(1+i) - D)dF(z) \text{ for } \gamma = 0, 1.$$

Next, the private insurer has no purchasing power with outside hospitals. Thus, we will normalize the private insurer's threat point to 0. So, with a $(r(\gamma), 0)$ threat point, the generalized Nash bargaining solution β^* maximizes $(Y - r(\gamma))^q V^{1-q}$ while guaranteeing V, $Y - r(\gamma) \geq 0$, where $q \in (0,1)$ is the relative bargaining power of the hospital.

3.3 Debt Strategies

Suppose that there are no costs to bankruptcy ($b = 0$) and that technology is reversible ($\gamma = 1$). In such a setting the hospital always chooses an all-equity capital structure. However, once we introduce bankruptcy costs ($b > 0$), the hospital's optimal capital structure will now involve debt.

Proposition 5 *When bankruptcy is costly ($b > 0$) and technology is reversible ($\gamma = 1$), the hospital issues a positive level of debt $D^* > 0$, such that the probability of bankruptcy is positive, $F(z^*) > 0$, in equilibrium. In some cases, the hospital may even repurchase some of its equity ($s^* < 0$) in order to leverage the hospital even*

further. The Nash bargaining solution β^* satisfies

$$Y = (1 + b\frac{\hat{X}(T)}{X(T)})V, \text{ where } \hat{X}(T) = \int_0^{z^*} X(T,z)dF(z).$$

Moreover, the rate β^* increases in debt D^*.

This debt strategy warrants special emphasis in our tax-free setting. Though we have focused on for-profit hospitals, this debt strategy can be shown to hold for nonprofit hospitals as well. In practice, since there is no applicable corporate income tax for not-for-profit hospitals, we would expect nonprofit hospitals to maintain an all-equity structure in order to avoid the risk of bankruptcy[8]. Yet both nonprofit and for-profit hospitals are highly leveraged relative to most other industries. Proposition 6 offers an interesting explanation of this phenomenon. First, since the private insurer enters price negotiations with the hospital after having observed the hospital's capital structure and investment level, the private insurer can behave opportunistically in the negotiations, as we have seen previously. To counter this opportunism, the hospital precommits to a highly levered position. This increases the risk of bankruptcy. Since bankruptcy is now costly, the hospital now demands a higher margin β. since both parties have equal bargaining power, the private insurer must concede to some increase in β. According to Proposition 6, the hospital's gain from this increase in the price per patient discharge outweighs its inflated risk of bankruptcy[9].

3.4 Conclusion

Though bankruptcy is rare in most regulated industries (e.g., public utilities), bankruptcy is not uncommon in the health care sector[10]. The reason is that the

[8]See Kraus and Litzenberger (1973), Scott (1976), and Flath and Knoeber (1981) for a discussion of the tax advantage of debt.

[9]Note that this result is very general in that we did not include a direct cost of bankruptcy in consumer utility V. If this were included in V, the private insurer would be even more apt to concede to an increase in β.

[10]Mullner (1983) documents a number of hospital closings. An average of 60 hospitals per year have closed since 1985, with about 100 closing in 1992. Closures greatly outnumber openings,

health care market is much more stochastic than other regulated markets. As a result, future demand is difficult to forecast. Our model indicates that hospitals are willing to use debt in the shadow of this highly volatile demand in order to counter the opportunism of the private insurance market.

and the overall supply of hospital beds declined by about 6 percent between 1980 and 1990 (AHA (1993)). Many financially distressed hospitals are not liquidated. As a typical example, the acute-care Doctor's Hospital of Tampa entered bankruptcy in 1992 and was reorganized as a long-term intensive care facility by Transitional Hospitals Corporation of Tampa.

CHAPTER 4
PRICING UNDER EXCLUSIVE DEALING

4.1 Introduction

Resale price maintenance (RPM) occurs when a manufacturer directly or indirectly limits the price at which its product can be sold by an independent retailer. *Minimum RPM* occurs when the manufacturer sets a *price floor* by enforcing only a minimum resale price, allowing the retailer to charge a higher price. *Maximum RPM* occurs when the manufacturer sets a *price ceiling.* Since the repeal of state Fair Trade laws in 1975, all forms of RPM have been *per se* illegal in the United States. However, many economists question this current *per se* illegality standard, arguing that some forms of RPM may generate efficiencies that benefit consumers. This debate has resulted in a large economic literature that analyzes the motives underlying RPM. Yet, this body of theory has failed to explain two fundamental observations from the RPM litigation cases that emerged with the repeal of state Fair Trade laws. First, from the Federal Trade Commission Report by Ippolito [1988] which examined all RPM litigation from 1976 to 1982, it is evident that both minimum RPM and maximum RPM are sometimes observed in the same industry across different markets[1]. Second, only exclusive dealerships were observed in the industries where both minimum RPM and maximum RPM were imposed. In contrast, *common* retailers were employed in almost all of the industries in which only maximum RPM appeared (i.e., rival manufacturers shared the same retailer[2]). This paper presents a model of manufacturer competition that explains these two findings concerning the connections between RPM and industry structure.

[1]For example, in 18 cases involving gasoline retailing, half involved allegations of price floors while the remainder involved price ceilings.

[2]Examples of common retailers in the 32 maximum RPM litigation cases include newspaper distributors, food and beverage distributors, and retailers of appliances, electronics, and machinery.

In most manufacturing scenarios, retailers are often better informed about market demand than manufacturers. Furthermore, a moral hazard problem is often present because demand is influenced by the (unobservable) marketing intensity and promotional efforts of the retailer. We construct a model characterized by asymmetric information and moral hazard and develop predictions consistent with the empirical findings of Ippolito. More precisely, we find that if the manufacturers compete through the same shared retailer, price ceilings emerge systematically for every realization of demand. In contrast, if the manufacturers instead compete head-to-head for the exclusive services of the retailer, price floors are optimal when demand is low (i.e., in weak markets) while price ceilings emerge when demand is high (i.e., in strong markets). Thus, the manufacturer's use of RPM depends critically on the industrial structure in a manner that is consistent with the empirical findings.

There are other interesting differences that emerge in the incentive contracts under these two modes of manufacturer competition. First, when manufacturers compete through the same retailer, the resale price and sales level that emerge in equilibrium will always differ as the level of the market demand differs. In contrast, when manufacturers compete head-to-head for the exclusive services of the retailer, a fixed, uniform retail price may emerge. In particular, the winning manufacturer institutes a rigid resale price that does not vary with demand in a range of intermediate demand levels. Thus, in a dynamic setting, an intertemporal price "stickiness" may emerge which is due entirely to healthy manufacturer competition. This finding stands in contrast to the traditional view that price rigidity is the result of tacit manufacturer collusion (Maskin and Tirole (1988) and Eaton and Engers (1990)[3]).

Rigid prices emerge in our model because head-to-head competition for an exclusive dealership creates countervailing marketing incentives for the retailer. The usual incentive for the retailer to understate demand arises to convince the manufacturer

[3]Prior to these two repeated-game models, price rigidities in oligopoly were first studied in the classical kinked-demand models of Hall and Hitch (1939), Sweezy (1939), and Stigler (1947). See Tirole (1988) for a general review of this literature.

that high sales are due to the retailer's diligent efforts, not exogenous demand. This incentive is countered by the following effect of competition. As market size increases, the total value of the retail outlet increases, leading to higher manufacturer bids for the retailer's services. The prospect of higher bids can encourage the retailer to exaggerate the size of the market. These countervailing incentives give rise to partial pooling in the equilibrium contracts. For weak markets, the incentive to exaggerate the market size prevails, and, to mitigate this incentive, price floors are imposed. For strong markets, the incentive to understate demand dominates, and, to help control this incentive, price ceilings are instituted. In the intermediate market range, these two incentives begin to conflict, as do the price ceilings and floors. The result of this conflict is to enforce the same uniform resale price regardless of demand, provided demand is of intermediate size. Consequently, in a dynamic setting, price adjustments will be more sluggish in the interim period between booms and busts than during booms or busts. This prediction differs from Maskin and Tirole's (1988) finding that price adjustments should occur least often during booms.

The traditional explanation for RPM under manufacturer competition is that manufacturers use RPM to sustain collusion (Telser (1960) and Posner (1977)). In contrast, we find that RPM is the result of intense manufacturer competition for a privately informed retailer. Although RPM under asymmetric information has been investigated in a setting with downstream retailer competition (e.g., Katz [1989] and Rey and Tirole [1986][4]), little attention has been afforded to RPM resulting from upstream manufacturer competition. An exception is Perry and Besanko [1991] who examine RPM under manufacturer competition for exclusive dealerships. They abstract from incentive conflicts and examine a restricted class of franchise contracts with RPM. They find that if manufacturers are restricted to use only fixed fees and/or linear wholesale prices with RPM, one equilibrium with maximum RPM arises, as

[4]In Rey and Tirole [1986], the retailers learn demand after contracting. In our model the retailer knows demand before contracting.

does a second Pareto-dominating equilibrium with minimum RPM. Thus, their model predicts that minimum RPM will appear systematically under competition for exclusive dealers. In contrast, our exclusive dealing game exhibits an undominated equilibrium that supports both minimum and maximum RPM in different markets at the same time [5] . This is consistent with the RPM pattern that Ippolito finds under exclusive dealing. Moreover, we do not restrict franchise contracts to any particular form.

According to Ippolito's second empirical finding, the systematic use of maximum RPM is observed only in industries where retailers are shared by rival manufacturers. We develop a common agency retailing model in the spirit of Stole (1992), Bernheim and Whinston (1992), and Martimort (1992,1993) that supports this second finding on RPM. Although these authors study the choice between exclusive dealing and common retailing under manufacturer competition in detail[6], they do not examine the important connection between RPM and the industry structure. For example, Bernheim and Whinston (1992) allow resale price to be an unobservable decision of the retailer, and analyze a moral hazard common agency problem. In contrast, we assume resale prices are observable, and examine an adverse selection common agency model in the spirit of Stole (1992) and Martimort (1992,1993). Their models of manufacturer competition consider retailers with private information about the costs of supplying the final good. The demand functions are common knowledge. In contrast, we allow the retailer to possess private information about market demand.

[5]The single-manufacturer analog of our model is developed by Blair and Lewis (1994). They find that both forms of RPM may emerge in a monopolized industry across different markets, but only when the standard Spence-Mirrlees sorting condition fails to hold. This sorting condition ensures that the retailer is compensated less when demand is high than when it is low for a given increase in the sales level or resale price. In our model we show that both forms of RPM may emerge even when this sorting condition is maintained. It is interesting to note that Romano [1994] has shown that both forms of RPM may emerge under complete information, depending on an expanded elasticity condition, if the manufacturer, as well as the retailer, must make an unobservable decision that affects demand (e.g., quality choice, national advertising, etc.).

[6]The choice of industry structure has also been studied by Besanko and Perry [1993], Aghion and Bolton [1987], Comanor and Frech [1985], Mathewson and Winter [1987], and Schwartz [1987].

As a result, our common agency model is multidimensional in the sense that both price and quantity must now be controlled by the manufacturer's contract since demand is not known *ex ante* by the manufacturers.

This chapter is organized as follows. In Section 2 we develop the common agency model. We demonstrate that the equilibrium contracts result in maximum RPM in every market when manufacturers compete through a common retailer. The equilibrium contracts are nonlinear: simple two-part wholesale price contracts with RPM cannot be sustained in equilibrium. In Section 3, we contrast the distortions induced in a common agency with those induced by a multiproduct monopolist. Section 4 delineates conditions under which both forms of RPM may emerge when manufacturers compete head-to-head for the exclusive services of the retailer. In addition, rigid retail prices are shown to emerge. Finally, Section 5 offers some concluding comments. The proofs of all the key findings are provided in the Appendix.

4.2 Common Agency

In this section we will examine the case of manufacturers competing through the same shared retailer[7]. Consider two vertically differentiated duopolists, M_1 and M_2, who wish to sell their products in the same market. The demand for both goods in this market can be parameterized by a scalar $\theta \in \Theta \equiv [\underline{\theta}, \overline{\theta}]$. Higher realizations of θ correspond to higher market demand in the sense described below. While θ is unknown to the manufacturers, it is common knowledge that θ has distribution $F(\theta)$. In contrast, a potential retailer knows the demand realization θ from the outset. The

[7]We are modelling a common agency with unknown demand. Recently, Stole [1992], Martimort [1992,1994], Ivaldi and Martimort [1993], Mezzetti [1993], and Gal-Or [1991a] have examined common agencies with unknown costs. Our common agency model is multidimensional in the sense that both price and quantity must now be controlled by the manufacturer's contract since demand is not known *ex ante* by the manufacturers. At the opposite extreme, Bernheim and Whinston [1985,1986,1992] and Fraysse[1993] have developed common agency models for the moral hazard case where all parties are *ex ante* uninformed about demand.

manufacturers are assumed to offer take it or leave it contracts simultaneously to the retailer.

On the demand side, we assume that the retailer cannot price discriminate. The retailer sets a single (observable) linear price for all customers. As a matter of exposition, we will first assume that the retailer's sales levels are observable. We show later that the manufacturers' optimal contracts would not change if sales could not be monitored, provided resale prices are observable. The demand for product i is given by $D_i(p_1, p_2, e_1, e_2, \theta)$, i=1,2. Demand for product i increases with the retailer's product-specific marketing intensity or promotional effort e_i. Moreover, due to positive promotional spillover effects, demand for product i may increase in e_{-i}. For a large class of demand functions and disutility functions $e(e_1, e_2)$, there will be a unique disutility minimizing effort pair (e_1, e_2) which will ensure that (x_1, x_2), and no more, is sold at the price vector (p_1, p_2) in a market of size θ[8]. Thus, instead of specifying the demand functions directly, it will be more convenient to work with the well-defined composite disutility function $e(x_1, x_2, p_1, p_2, \theta)$, which is the minimum cost or disutility incurred by the agent in insuring that (x_1, x_2), and no more, is sold at the price vector (p_1, p_2) in a market of size θ. As θ increases, the level of effort needed to sell (x_1, x_2) at price (p_1, p_2) declines. The retailer's effort may include marketing intensity, advertising, or such customer services as free delivery and installation, free repair or consultation services, and product demonstration. This type of effort is not readily observed by the manufacturers, so a moral hazard problem arises. That is, the manufacturers cannot verify directly whether low demand occurs because the retailer has supplied little effort or because demand is truly sluggish.

On the supply side, manufacturer M_i produces x_i at constant marginal cost m_i, and imposes a tax $A_i(x_i, p_i)$ on the retailer. $A_i(x_i, p_i)$ is the amount the retailer must pay to manufacturer M_i when he chooses to sell x_i units of M_i's product at price p_i.

[8]For example, consider the demand functions $D_i(p_1, p_2, e_1, e_2, \theta) = d_i(p_1, p_2, \theta) + b_{i1}e_1 + b_{i2}e_2$, with $b_{11}b_{22} - b_{12}b_{21} > 0$ and $e(e_1, e_2)$ convex.

Both pricing and sales decisions are delegated to the agent. Although price is not directly dictated by the manufacturer, she can indirectly control the price through the tax since prices are observable. Thus, manufacturer M_i seeks to solve the following program in the common agency duopoly game:[9]

$$[\mathrm{CA}] \qquad \text{maximize}_{A_i(\cdot,\cdot)} \int_{\Theta} \{A_i(x_i(\theta), p_i(\theta)) - m_i x_i(\theta)\} f(\theta) d\theta,$$

where the multi-outlay (x_1, x_2, p_1, p_2) is chosen by the retailer to solve

$$\max_{x_1, x_2, p_1, p_2} \{x_1 p_1 + x_2 p_2 - e(x_1, x_2, p_1, p_2, \theta) - A_1(x_1, p_1) - A_2(x_2, p_2)\}.$$

The timing in the game is as follows. First, the agent learns the realization of θ. Next, the manufacturers simultaneously offer the agent their tax schedules. The agent then accepts either both, one, or none of the tax schedules. If the retailer rejects both contracts, he earns his reservation wage, which is normalized to zero. In this section, we will derive a perfect Bayesian equilibrium of this two stage game.

Given the complexity of the game, a characterization of the equilibrium is provided for the case where the retailer's disutility of effort is quadratic, and takes the general form:

$$\begin{aligned} e(x_1, x_2, p_1, p_2, \theta) \equiv \frac{a_1}{2} x_1^2 + \frac{a_2}{2} x_2^2 + \frac{b_1}{2} p_1^2 + \frac{b_2}{2} p_2^2 + h_1 x_1 p_1 + \\ + h_2 x_2 p_2 - c x_1 x_2 - c p_1 p_2 - \theta(x_1 + x_2 + p_1 + p_2). \end{aligned} \tag{4.1}$$

Notice that the form in (1) admits an approximation of many disutility functions up to a second-order Taylor expansion. We will assume that a_i and b_i are positive and large enough that the retailer's disutility of effort must increase as the price or the required sales level increases, i.e., $e_{x_i}, e_{p_i} > 0$ for i=1,2. Also, $a_i b_i - (1 - h_i)^2 > 0$ is a necessary condition for the retailer's program in [CA] to be concave in equilibrium.

[9] We do not allow manufacturer M_i to offer contracts that are contingent upon the agent's choice of x_{-i} and p_{-i} for the rival's product. Such contingency contracts are rarely observed in practice since they are difficult to enforce legally. Bernheim and Whinston [1992] examine contingency contracts in a moral hazard framework. They find that contingency contracts may result in the existence of economically implausible equilibria.

In addition, we will require that $h_i \leq 1$ for i=1,2. This natural assumption ensures that as the resale price increases, each unit of sales becomes more profitable for the manufacturer under complete information. Finally, note that the two product lines will be substitutes if $c < 0$ and complements if $c > 0$.

When the manufacturers have complete information on product demand, the equilibrium contract is extremely simple, consisting solely of a fixed franchise fee. This fee extracts all the rents from the retailer. But, more importantly, RPM does not occur under complete information. However, if the retailer possesses private information on product demand, the equilibrium franchise contracts become more complex and involve RPM, as the next Theorem illustrates[10]. First, we will make use of the following assumption throughout the paper.

Assumption 3 θ has an increasing affine hazard rate. [11].

Due to the highly nonlinear nature of e, it is extremely difficult to develop an algorithm to compute the equilibria when the hazard rate of the distribution is also nonlinear. Under Assumption 1 we have the following existence and characterization result.

Theorem 1 There exists a pure-strategy equilibrium for the common agency game [CA]. In this equilibrium, franchise taxes take the form

$$A_i(x_i, p_i) = \gamma_i + \alpha_i x_i + \frac{\beta_i}{2} x_i^2 + \delta_i p_i + \frac{\eta_i}{2} p_i^2 + \sigma_i x_i p_i \tag{4.2}$$

for every $(x_i, p_i) \in \Re_+^2$, i=1,2. This equilibrium schedule in (2) is fully separating in that it always induces the retailer to select an outlay (x_i, p_i) that differs strictly with

[10]Due to the moral hazard problem, the first-best contract cannot be implemented as in the asymmetric demand information models of Lewis and Sappington [1988], [1992].

[11]An increasing hazard rate $\frac{f(\theta)}{1-F(\theta)}$ helps to ensure that the tax will be fully separating. The hazard rate is affine if θ is generated by a Beta$(1, \frac{1}{\lambda})$ density on Θ with distribution $1 - \frac{(\overline{\theta}-\theta)^{\frac{1}{\lambda}}}{(\overline{\theta}-\underline{\theta})^{\frac{1}{\lambda}}}$ for $\lambda \geq 0$. Ivaldi and Martimort [1993] use these Beta distributions to simplify the analysis of duopolistic price discrimination when θ is bivariate. Note that as λ increases, the density shifts more weight to the higher realizations of demand. In fact, as $\lambda \to \infty$, the distribution approaches the Dirac mass concentrated at $\overline{\theta}$. Note that $\lambda = 1$ is the uniform density.

the realized market size θ. Furthermore, the equilibrium taxes entail a fixed franchise fee (γ), a nonlinear wholesale price (α, β), a nonlinear tax on the retail price (δ, η), and a royalty (σ) on gross revenue. Equilibrium royalty rates are $\sigma_i = 1 - h_i$, i=1,2. There does not exist any simpler combination of a two-part wholesale price and/or royalty on revenue in equilibrium.

The retailer's equilibrium profits increase with demand, θ. When the products are substitutes ($c < 0$), the retailer earns rents even in the weakest market $\underline{\theta}$. Under complements, the retailer earns only his reservation wage in the weakest market $\underline{\theta}$.

Note that the resale price is never *directly* imposed on the retailer, but it is controlled indirectly. Resale price maintenance appears as a levy on resale price and a royalty on revenue. [12].

To understand the equilibrium pattern of rents, first consider the case where the two products are substitutes. In this case, the retailer naturally has an incentive to serve one manufacturer exclusively. Thus, the retailer has to be afforded extra participation rents just to agree to serve *both* manufacturers. So, to induce participation, the retailer is allowed to retain a premium rent even in the weakest market $\underline{\theta}$. Under complementary products, the agent does not have to be forced to serve as a common agent, and so the retailer can be forced to zero rents in the weakest market[13].

To understand the nature of the royalties on revenue (σ_i), recall that h_i is the degree of substitution between sales x_i and resale price p_i in the promotional cost function or marketing function e. Since $\sigma_i = 1 - h_i$, the royalty on product i's revenue increases at the same rate that the degree of promotional substitution between the resale price and the quantity of good i decreases. In particular, if resale price and

[12]Although taxes on the final resale price seem rare in practice, they have been advocated in the regulation literature. Laffont and Tirole [1990] and Leung [1994] have proposed using a tax on price to regulate a natural monopoly. Even though A_i is concave, it is still an open question of whether there exists an equivalent equilibrium in which both retailers offer a menu of affine contracts of the form $\mathcal{A}_i(x_i, p_i, \theta) = \xi_0(\theta) + \xi_1(\theta)x_i + \xi_2(\theta)p_i$.

[13]This results in a continuum of Nash equilibria under complements since there is a continuum of (γ_1, γ_2) equilibrium pairs that satisfy the reservation wage constraint for any fixed set of equilibrium parameters $(\alpha_i, \beta_i, \delta_i, \eta_i, \sigma_i)$, i=1,2. Under substitutes, this corresponding pair is unique.

quantity are promotional cost complements of degree $h_i = 1$, then no royalty is imposed on revenue. Note that this equilibrium royalty on revenue extracts all the revenue from the agent's sale of product i less the promotional cost term $h_i x_i p_i$ of e. Consequently, the fixed franchise fee, the nonlinear wholesale price, and the nonlinear surcharge on resale price must act as a combined subsidy to the retailer to recoup the remainder of his promotional costs and to provide any necessary rents.

Next, we examine the more general case in which the manufacturers cannot observe the retailer's final level of sales. Below, $A_i(x_i, p_i)$, i=1,2, will refer to an equilibrium of the common agency game [CA] in Theorem 1, where sales were monitored.

Corollary 3 *Suppose the manufacturers cannot observe the retailer's level of sales, so the retailer can sell any q_i less than the ordered inventory x_i, i=1,2. Then $A_i(x_i, p_i)$, i=1,2, will still induce the retailer to sell the exact quantity that he orders, i.e., $q_i = x_i$. Moreover, the equilibrium franchise taxes $A_i(x_i, p_i)$, i=1,2, indirectly impose price ceilings and quantity rationing on the retailer in every market θ.*

That is, $A_i(x_i, p_i)$, i=1,2, persists as an equilibrium franchise tax for the more general duopoly game where manufacturers do not monitor the retailer's sales levels. In addition, $A_i(x_i, p_i)$ does not give the retailer the incentive to destroy or to store excess goods. In fact, the franchise taxes induce the retailer to choose resale prices and sales levels that are lower than what he would choose if he were the full residual claimant[14]. Thus, as a result of manufacturers competing through a shared retailer, the franchise tax indirectly imposes maximum, not minimum, RPM for every realization of demand. This is in accord with Ippolito's examination of RPM litigation cases. In almost all the cases from 1976 to 1982 involving allegations of only maximum RPM, the retailer was shared by rival manufacturers.

[14]The retailer would be the full residual claimant is if each manufacturer M_i offered a pooling contract consisting of only a wholesale price of m_i.

4.3 The Monopoly Benchmark

As a benchmark, we now examine a monopolist who markets both x_1 and x_2 through a single retailer when there is no threat of entry by a second manufacturer. To facilitate comparison with the above duopoly case, we make the following definitions. Let $p_i^D(\theta)$ and $x_i^D(\theta)$, i=1,2, be equilibrium resale prices and sales levels chosen by the retailer in the common agency (duopoly) game [CA]. Let $p_i^M(\theta)$ and $x_i^M(\theta)$, i=1,2, be the optimal resale prices and sales levels chosen by the retailer serving a multiproduct monopolist. Define $p_i^F(\theta)$ and $x_i^F(\theta)$, i=1,2, to be the corresponding (first-best) levels for a monopolist who has perfect knowledge of demand. Similarly, define the equilibrium disutility of effort levels e^D, e^M, and e^F accordingly. We can now examine the efficiency of duopoly competition through a common retailer in the game [CA].

<u>*Proposition 6*</u> *Total manufacturer profits are higher under monopoly than under the duopoly of game [CA]. Furthermore:*

1. *For substitute products ($c < 0$), $x_i^F(\theta) \geq x_i^D(\theta) \geq x_i^M(\theta)$ and $p_i^F(\theta) \geq p_i^D(\theta) \geq p_i^M(\theta)$, i=1,2, with equality only for the largest market $\overline{\theta}$. The price ceiling and quantity rationing are less severe under duopoly. The retailer retains more rents under duopoly than under monopoly and works more efficiently: $e^F(\theta) \geq e^D(\theta) \geq e^M(\theta)$, with equality only for the largest market $\overline{\theta}$.*

2. *For complementary products ($c > 0$), $x_i^D(\theta) \leq x_i^M(\theta) \leq x_i^F(\theta)$ and $p_i^D(\theta) \leq p_i^M(\theta) \leq p_i^F(\theta)$, i=1,2, with equality only for the largest market $\overline{\theta}$. The price ceiling and quantity rationing are more severe under duopoly. The retailer secures less rents under duopoly than under monopoly and works less efficiently: $e^F(\theta) \geq e^M(\theta) \geq e^D(\theta)$, with equality only for the largest market $\overline{\theta}$.*

Thus, for substitutes, the duopoly price ceilings of game [CA] are higher than the monopoly price ceilings. However, quantity rationing is less severe under competition

(see Figure 3). Although a duopoly provides higher-powered incentives to the retailer than a monopoly would under substitutes, the total manufacturer profits are still higher for the multiproduct monopoly than for the duopoly. To understand why, recall that for substitutes, the retailer has a natural incentive to exclusively serve only one manufacturer. Thus, he must be provided with extra rents as an incentive to serve as a common retailer for both manufacturers. In contrast, under a monopoly the retailer does not have this outside opportunity, so the monopolist does not have to afford the retailer these extra duopolistic participation rents.

Price ceilings persist in the duopoly common agency game [CA] when products are complements. Furthermore, competition induces lower price ceilings and more severe rationing. Just as a monopolist will, the duopolists distort price and quantity downward as the market gets weaker in order to limit the retailer's incentives to understate demand. But, because the products are complements, if one duopolist decreases her resale price and level of sales, it becomes advantageous for the rival to also lower her price and sales level. As a result, the manufacturer's trade-off between efficiency and rent extraction becomes less severe as the rival reduces her price and sales level in order to extract rents. In essence, each manufacturer imposes a rent extracting externality upon the other when they share the same retailer, resulting in a double extraction of rents. Consequently, the retailer prefers a monopoly when the two product lines are complements.

When the goods are substitutes, if one duopolist reduces her price and sales level, the rival will find it beneficial to increase both sales and price. Thus, the duopolist's trade-off between efficiency and rent reduction becomes more severe as the rival reduces her price and level of sales in order to extract rents. Because of this externality under substitutes, the duopolist extracts less rents than the monopolist. As a result, incentives under a duopoly are higher-powered than the monopolist's incentives.

Finally, the underprovision of effort is worse under duopoly when the products are complements. Since a duopoly results in a lower level of sales at a lower resale price when compared with a multiproduct monopoly under complements, the level of effort required in a duopoly will be lower, and, hence, further from the first best level. The exact opposite occurs for substitute product lines; duopoly competition will induce the retailer to provide a higher, more efficient level of marketing intensity.

4.4 Competition for an Exclusive Dealer

In the last section we saw that when products are substitutes the retailer had to be afforded extra rents just to agree to serve *both* manufacturers. Eventually, these participation rents can become so costly for a manufacturer that her expected profits are higher if she instead hires the retailer as an exclusive dealer. To consider this possibility, we now analyze the case where, instead of sharing the same agent non-cooperatively, the manufacturers compete head-to-head for the exclusive services of the agent.

The salient feature of this model is that head-to-head competition for exclusive services creates countervailing marketing incentives for the retailer[15]. The usual incentive for the retailer to understate demand arises to convince the manufacturer that high sales are due to the retailer's diligent efforts, not exogenous demand. This incentive is countered by the following effect of competition. As market size increases, the total value of the retail outlet increases. Hence, the manufacturers' bids for the

[15] For a general discussion of countervailing incentives in principal-agent problems, see Lewis and Sappington [1989a,b]. Our model of exclusive agency differs significantly from the model of Biglaiser and Mezzetti [1993]. They examine the optimal labor contract that emerges when two firms compete for a manager whose ability is hidden. Their model is structured so that countervailing incentives do not arise. Our model of exclusive dealing also differs from Gal-Or (1991b) and Martimort (1993). They consider two manufacturers, each of whom operates through an exclusive dealer. The manufacturers compete through wholesale prices, leaving the resale price to be determined by downstream competition (Gal-Or also considers the case of RPM). Instead, we follow Bernheim and Whinston (1992) in considering two manufacturers who compete for the exclusive services of *one* retailer.

retailer's services increase with the perceived market size. The prospect of higher bids can encourage the retailer to exaggerate the size of the market. These countervailing incentives will give rise to partial pooling in the equilibrium contracts.

In this competitive scenario, each manufacturer offers the retailer a menu of contracts $(x_i(\theta), p_i(\theta), T_i(\theta))$ in exchange for his exclusive services. $T_i(\theta)$ is a fixed fee. No wholesale price is necessary since now x_i is dictated directly. Once the agent accepts M_i's contract, he is prohibited from dealing with manufacturer M_{-i}, who then has no access to this particular market. Since this is a static model, we assume that a manufacturer can never credibly bid more than the maximum total surplus value of her product. That is, the most that M_i can bid to attract the retailer is

$$\Pi_i(\theta) \equiv (p_i^*(\theta) - m_i)x_i^*(\theta) - e(x_i^*(\theta), 0, p_i^*(\theta), 0, \theta),$$

where (x_i^*, p_i^*) is the outlay that the retailer would choose if the manufacturer's exclusive contract consisted solely of a wholesale price equal to marginal cost m_i. Thus, $\Pi_i(\theta)$ is the profit that the retailer would obtain if he were the full residual claimant of product line i. Equivalently, $\Pi_i(\theta)$ is the vertically integrated profit for M_i. Note that the total surplus $\Pi_i(\theta)$ increases as the market demand θ increases. For simplicity, we will assume that manufacturer M_1 dominates M_2 in the sense that $\Pi_1(\theta) \geq \Pi_2(\theta)$ for every θ. This will be the case, for example, if the two product lines are homogeneous and M_1 has a lower marginal cost than M_2. Alternatively, M_1 could dominate M_2 if product 1 possesses a stronger brand loyalty than M_2's product.

Since M_1 dominates, we will observe that the retailer always chooses to exclusively serve M_1 in equilibrium. Thus, an equilibrium strategy for M_2 is simply to offer the pooling contract consisting of a wholesale price equal to marginal cost m_2, while M_1's

equilibrium strategy (x_1, p_1, T_1) solves the following restricted program[16]:

$$[ED]\begin{cases} \max_{p_1,x_1,T_1} \int_{\Theta} \{T_1(z) - m_1 x_1(z)\} f(z) dz \text{ such that} \\ (1)\ \ U'(\theta) = x_1(\theta) + p_1(\theta), \\ (2)\ \ U(\theta) \geq \Pi_2(\theta), \text{ and} \\ (3)\ \ p_1 \text{ and } x_1 \text{ are nondecreasing in } \theta, \end{cases}$$

where

$$U(\theta) = U(\theta,\theta) \text{ and } U(\tilde{\theta},\theta) \equiv x_1(\tilde{\theta})p_1(\tilde{\theta}) - e(x_1(\tilde{\theta}), 0, p_1(\tilde{\theta}), 0, \theta) - T_1(\tilde{\theta}).$$

To facilitate a comparison with the duopoly model, we will continue to assume that e has the general quadratic form given in equation (1). In addition to Assumption 1, we will restrict attention to those Beta distributions with $\frac{F(\theta)}{f(\theta)}$ nondecreasing in θ. An example is the uniform distribution ($\lambda = 1$). With these assumptions we have the following characterization of the optimal vertical restraints under competition for an exclusive dealer.

<u>Proposition 7</u> *The participation constraint (2) in program [ED] is binding at a unique point $\hat{\theta} \in (\overline{\theta}, \underline{\theta})$. Moreover, at the solution (x_1, p_1) to [ED] we have the following* [17]*:*

1. *For sluggish markets ($\theta < \hat{\theta}$), the optimal vertical restraint involves resale price floors and quantity forcing. Also, $p_1(\theta) \geq p_1^*(\theta)$ and $x_1(\theta) \geq x_1^*(\theta)$ for $\theta \leq \hat{\theta}$,*

[16]We assume that M_1 wins any ties, that Π_i is concave in x_i and p_i, and that, in equilibrium, the agent will be employed for any realization of demand. In addition, it is important to understand how program [ED] is restricted. Note that the optima under the more general program with the constraint $U(\theta,\theta) \geq U(\tilde{\theta},\theta)$ (instead of constraints (1) and (3)) are not always nondecreasing. However, in most cases, these optima are nondecreasing so that program [ED] is indeed the correct program.

[17]Since $\Pi_2(\theta)$ is increasing in θ, it is not immediately clear for which θ's the participation constraint (2) will be binding. However, due to the strict convexity of $U(\theta) - \Pi_2(\theta)$, we can conclude that constraint (2) binds only at one point, $\hat{\theta}$. We assume that $\Pi'_2(\theta) \in (x_1^H(\underline{\theta}) + p_1^H(\underline{\theta}), x_1^L(\overline{\theta}) + p_1^L(\overline{\theta}))$ for all $\theta \in [\underline{\theta}, \overline{\theta}]$, so that $\hat{\theta} \in (\underline{\theta}, \overline{\theta})$, where $(x_1^L(\theta), p_1^L(\theta))$ solves program [ED] with constraint (1) modified as $U'(\theta) = x_1(\theta) + p_1(\theta) - \Delta$, where Δ is sufficiently large so that constraint (2) binds only at $\overline{\theta}$. In contrast, $(x_1^H(\theta), p_1^H(\theta))$ solves program [ED] when $\Delta = 0$, so that (2) binds only at $\underline{\theta}$. That is, $(x_1^H(\theta), p_1^H(\theta))$ is the optimal contract when M_2 does not threaten to enter the market, i.e., $\Pi_2 = 0$. Thus, $(x_1^H(\theta), p_1^H(\theta))$ is the optimal franchise menu for Blair and Lewis' [1994] model. Moreover, it is the optimal menu for the common agency model of Section 2 when c=0.

with equality only at $\underline{\theta}$ and $\hat{\theta}$. The retailer's informational rents decrease in θ with zero rents at $\hat{\theta}$ (wages at $\hat{\theta}$ are $\Pi_2(\hat{\theta})$). Moreover, there is an overprovision of promotional effort, with efficient effort supplied only at $\underline{\theta}$ and $\hat{\theta}$.

2. *For strong markets ($\theta > \hat{\theta}$), the optimal vertical restraint entails resale price ceilings and quantity rationing. Also, $p_1(\theta) \leq p_1^*(\theta)$ and $x_1(\theta) \leq x_1^*(\theta)$ for $\theta \geq \hat{\theta}$, with equality only at $\overline{\theta}$ and $\hat{\theta}$. The retailer's informational rents increase in θ with zero rents at $\hat{\theta}$. Moreover, there is an underprovision of marketing effort, with efficient effort supplied only at $\overline{\theta}$ and $\hat{\theta}$.*

Quantity forcing requires that the manufacturer be able to monitor the retailer's level of sales. If such monitoring is not possible, the following adjustments in the optimal vertical restraints arise.

Corollary 4 *If the manufacturer cannot monitor the retailer's level of sales, the manufacturer elevates the price floor. The retailer is then induced to provide the efficient level of sales for the accompanying inflated price floor.*

Proposition 7 and Corollary 4 are illustrated in Figures 1 and 2. Three conclusions reported in Proposition 7 warrant special emphasis. First, price floors can be optimal vertical restraints in sluggish to intermediate-sized markets, regardless of whether sales are monitored. Due to the head-to-head competition between the manufacturers, the retailer's reservation utility increases with the market size, giving rise to incentives for the retailer to exaggerate the strength of the market. To mitigate this incentive, the manufacturer calls upon the retailer to generate performance that is particularly difficult to achieve if demand is below its reported level. The elevated price floor forces the retailer to put forth an above-normal level of marketing intensity to promote demand at this artificially high price, particularly if demand is low. Consequently, in a weak market, the price floor mitigates the retailer's incentive to exaggerate the market size to the manufacturer.

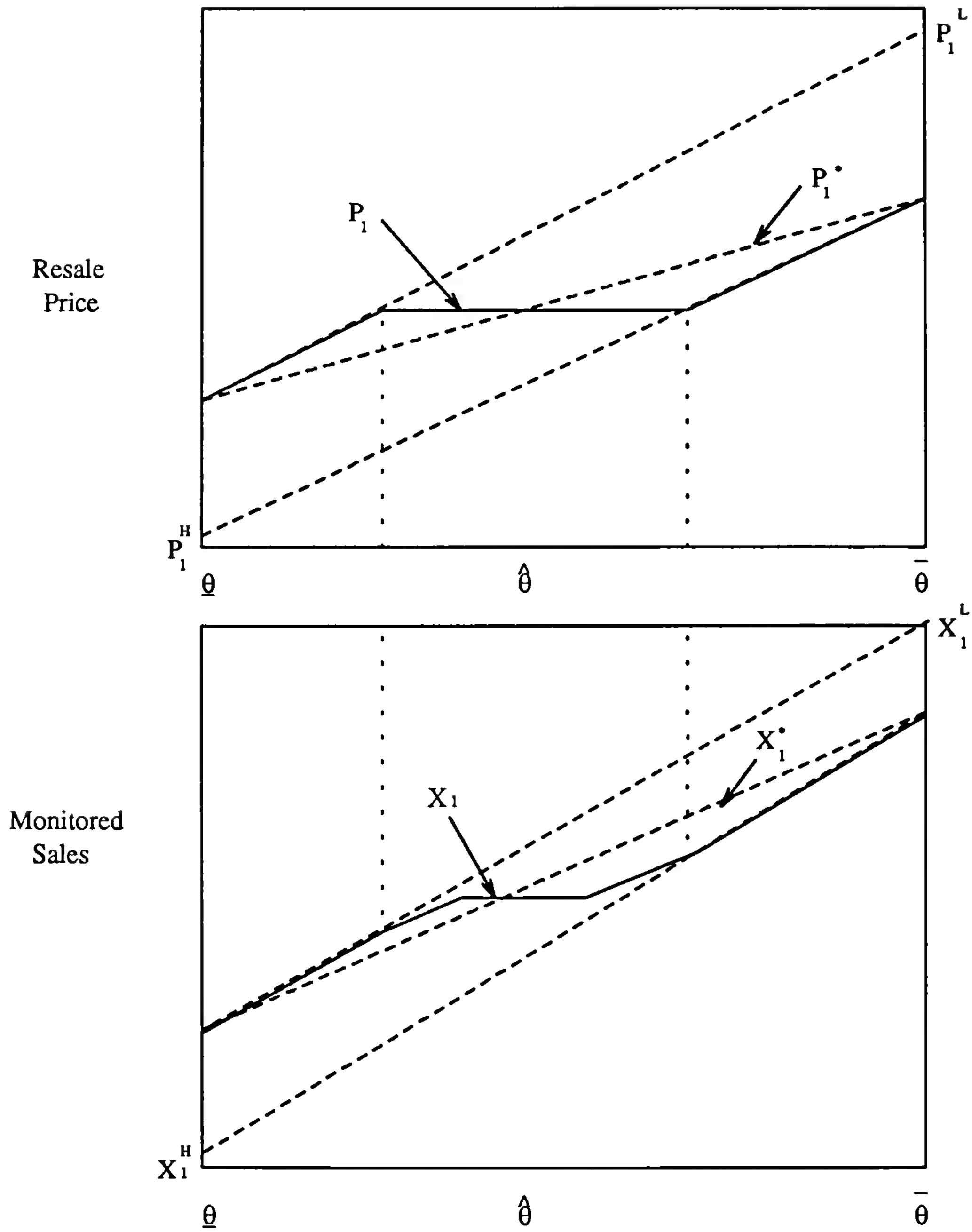

FIGURE 4-1 : MONITORED SALES.

Second, for an industry in which only exclusive dealers are employed, both price ceilings and price floors should emerge across different markets. We saw above that price floors should appear in weak markets. In contrast, price ceilings arise to limit the retailer's incentive to understate the size of strong markets. Recall that this was the prevailing incentive in the common agency model of Section 2. Thus, as in a common agency, this incentive is mitigated by the use of price ceilings and quantity rationing. In summary, we should thus observe both price ceilings and price floors across different markets. This prediction is consistent with the empirical findings of Ippolito [1988]. For example, in 18 resale price maintenance cases involving gasoline retailing, half involved allegations of price floors while the remainder involved price ceilings.

Third, rigid retail prices may emerge. For intermediate-sized markets, it is optimal for the manufacturer to institute a single, fixed resale price. Recall that the head-to-head competition between the manufacturers creates countervailing incentives for the retailer. The intermediate-ranged markets in which pooling occurs are essentially those markets in which the retailer has a "conflict" between the incentive to overstate and understate the size of the market θ. For sluggish markets and booming markets, the price mechanism is sensitive to market demand, continuously increasing in θ. However, the optimal retail price does not vary with the size of the market for intermediate-sized markets.

This price rigidity can be interpreted as an intertemporal price stickiness. In a dynamic setting, each manufacturer would prefer to commit to the static contract (derived above) in each period (see Laffont and Tirole (1990)). That is, they would prefer to offer the same static contract each period if they could credibly commit to this without renegotiating[18]. In such a case we would observe intertemporal price rigidities in intermediate markets even when demand is non-stationary. Consequently,

[18] It is interesting to note that empirical evidence indicates that franchise contracts rarely change over time. Banerji and Simon (1992) and Lafontaine (1993) find that often the same exact contract is offered each period to an exclusive dealer.

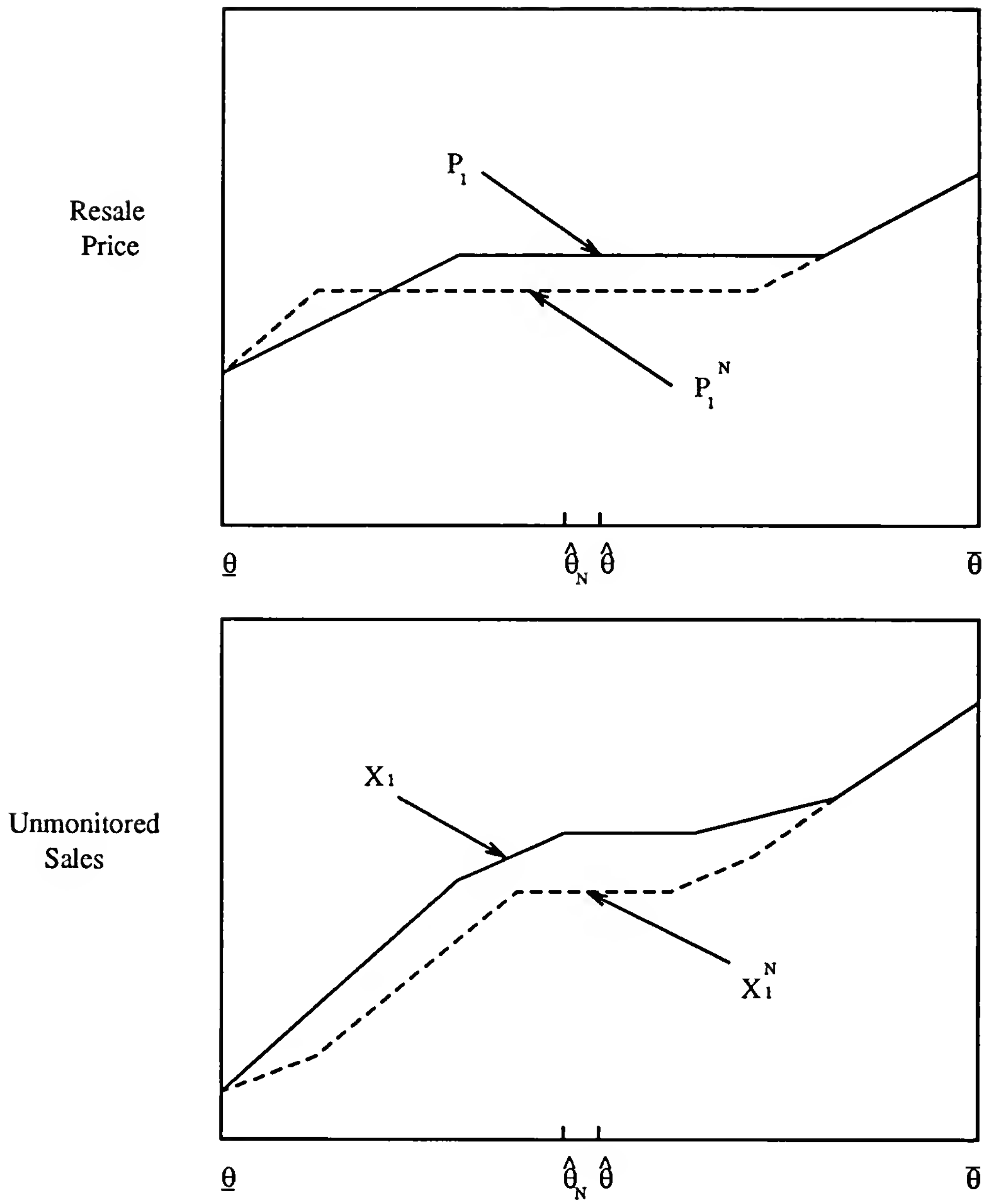

FIGURE 4-2 : UNMONITORED SALES

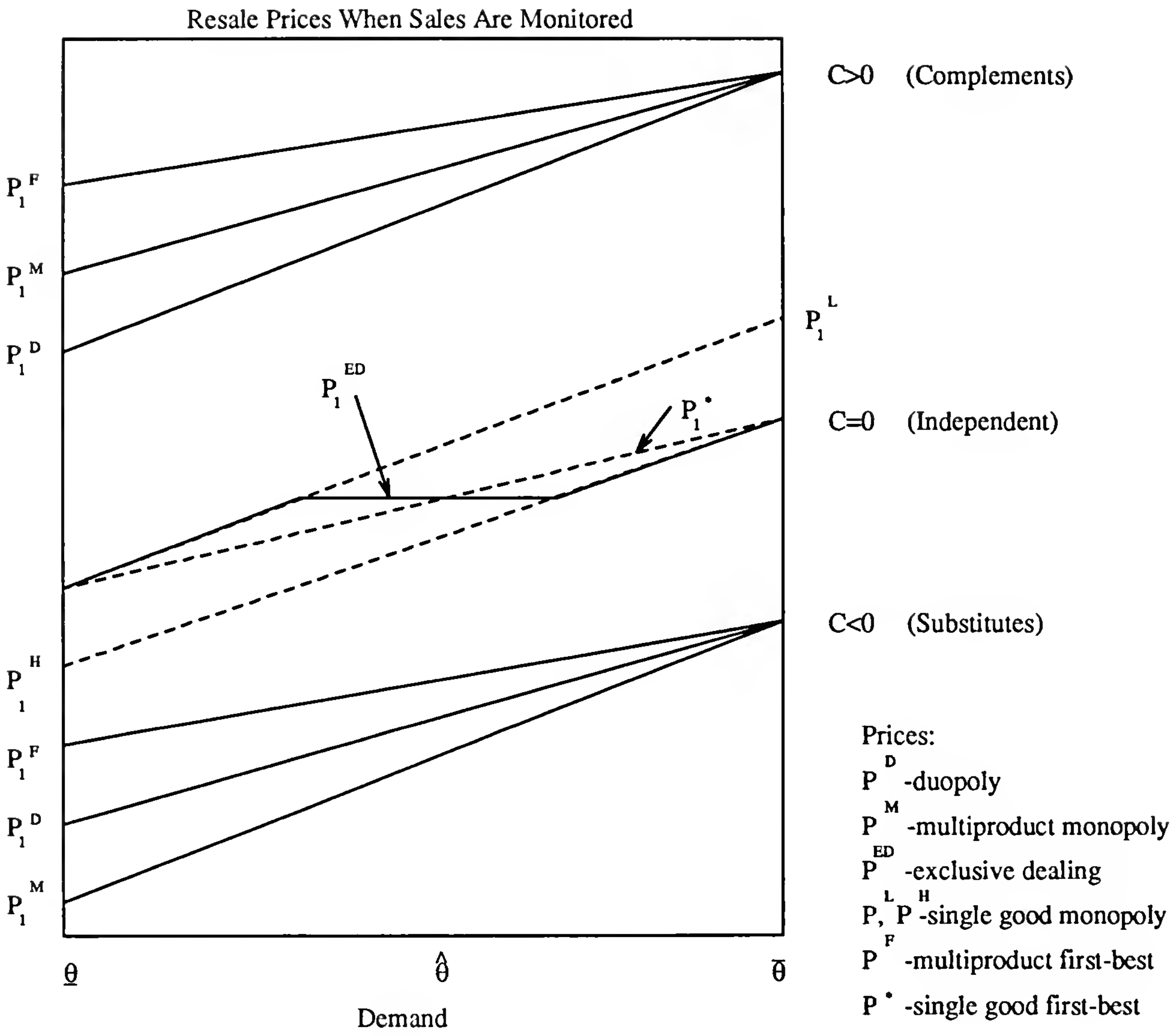

FIGURE 4-3 : COMMON AGENCY VS. EXCLUSIVE AGENCY

across time and across markets of considerably different sizes, we may observe the dominant manufacturer instituting the same exact retail price! There is no price discrimination. Note that this price stickiness does not arise out of risk-sharing concerns. The price rigidity is solely the consequence of incentive contracting under manufacturer competition. Finally, note that this constant price approaches the perfectly competitive price of m_1 as $\Pi_1(\theta) \rightarrow \Pi_2(\theta)$ for all pooling θ's, if the two products are homogeneous.

These results are due to intense manufacturer competition. In conclusion, we present Figure 3 to compare the vertical restraints that emerge under exclusive dealing with those that occur under duopoly. Figure 3 illustrates the predominant case where exclusive dealing results in a lower (higher) resale price when compared with the common agency retailing of complementary (substitute) goods.

4.5 Hospital Exclusive Dealing

The above common agency and exclusive dealing models can be reinterpreted to explain hospital competition rather than retailing in general. Consider the manufacturers to be hospitals offering a facility (such as an open heart surgery facility) to a physician group (the agent or retailer). The hospital M_i charges the physician group $A_i(x_i, p_i)$ for the use of the hospital facility. Here x_i is the number of patients that the physicians treat at hospital M_i. The price p_i is the fee the physicians charge each of those patients. Moreover, following Chapter 2, physicians can induce patient demand with their unobservable effort, intensity, or quality e_i. The physicians know the healthiness of the patient pool (HMO) that they serve. This risk is indexed by θ. High θ is an unhealthy pool of older patients, while lower θ is a pool of young, healthy patients. In contrast, the hospitals do not know the HMO's risk factor θ, and so must design incentive contracts to induce the physicians to reveal this information and to induce the desired level of hospital admissions. Each hospital offers a reimbursement rate to the physician group for the use of the hospital's technology.

The physician group than bundles this hospital rate with a physician reimbursement rate and charges this price bundle to the HMO.

In particular, when the two hospitals have complementary technologies, they will share the same physician group. The resulting price bundle will offer a discount to the HMO. When hospital technologies are substitutes, the dominant hospital will win an exclusive contract with the physician group. The resulting HMO price bundle will now involve "balanced billing" for low risk HMOs (e.g., HMOs with young, healthy enrollees). That is, the low risk HMO will be over-billed beyond what a vertically integrated hospital would charge. However, high risk HMOs will be offered a discounted price. Intermediate risk HMOs will be offered a uniform, risk-invariant price bundle. This competitive model may help explain many pricing structures observed in the health care industry.

4.6 Conclusion

In this paper we have derived equilibrium retail contracts for two rival manufacturers competing through a shared retailer who is privately informed about market demand. This scenario is very common in many industries. The equilibrium franchise taxes we identified are nonlinear, but they can be implemented by a menu of linear franchise contracts. The duopoly equilibrium has the following additional characteristics. For substitute (complementary) products, duopolistic manufacturer competition through a common retailer results in higher (lower) price ceilings, more (less) efficient underprovision of retailer promotional effort, and less (more) severe quantity rationing when compared with a multiproduct monopolist. In contrast, competition for the exclusive services of the agent results in price floors and an overprovision of effort in weak markets. This upward distortion was shown to be due to the presence of countervailing incentives. For strong markets, maximum RPM is still imposed.

Our model predicts that in industries where the rival manufacturers share a common retailer (e.g., newspaper distributors, food and beverage distributors, clothing distributors, etc.), only maximum RPM should arise. In contrast, in industries where retailers are exclusive dealers, both maximum and minimum RPM may arise simultaneously in different markets. These are precisely the patterns observed in Ippolito's (1988) study of RPM litigation cases.

An additional consequence of manufacturer competition for an exclusive dealer is that the same uniform resale price may be instituted in markets that may differ significantly in size. This, in turn, may lead to an intertemporal price stickiness. The price rigidity emerges in our model because of intense head-to-head competition between manufacturers, not because of tacit collusion among manufacturers, as other authors have suggested.

A direction for future research would be to investigate the effect that a ban on RPM would have on equilibrium contracts. In particular, it would be interesting to determine whether the final market structure (exclusive vs. common retailer) is affected systematically by a ban on RPM. Moreover, if sticky resale prices do not emerge systematically with a ban on RPM, then our model suggests that resale price rigidity (across markets and across time) may indicate that resale price maintenance is being practiced.

CHAPTER 5
CONCLUDING REMARKS

Common agency theory deals with a competitive version of principal-agent theory in which multiple principals contract with the same agent. This dissertation develops common agency theory for applications to the industrial organization of the health care industry.

In the first essay, the efficiency of three health care systems is examined. In a multi-payer system the public payer (Medicare) uses a mix of prospective payments and pass-through payments, while the private payer (a managed care insurer) uses a quality-based reimbursement rate through utilization review. Cost-shifting in the multi-payer system induces the hospital to overinvest in technology. Furthermore, pass-through payments of capital are scaled back in equilibrium since they create a moral hazard problem that allows the hospital to goldplate, i.e, to invest in wasteful, non-technological capital. This may explain Medicare's current policy to phase-out pass-through payments. In a single-payer system investment in technology is curtailed, and goldplating can arise. When trilateral negotiations between the payers and the hospital is admitted, technological efficiency results when a uniform quality-based reimbursement rate is negotiated. Moreover, this all-payer system is immune to goldplating.

The second essay studies the interrelationship between a hospital's capital structure and the payment plans designed by Medicare and the private insurance sector to reimburse the hospital's cost of capital and technology. To counter the opportunism of a managed care private insurance sector engaging in utilization review, the hospital will use debt when bankruptcy is costly.

The final essay derives the equilibrium hospital contract with a physician group that has private information on the risk rating of the health maintenance organization

it serves. Bundled physician charges are then derived when hospitals compete for an exclusive contract.

Finally, we suggest some theoretical extensions to our model. First, we have considered only deterministic patient demand in the first essay. In reality, the demand for medical services is highly stochastic. In such a case, some range of pass-through payments may be optimal if the hospital is risk averse. Moreover, as shown in the second essay's stochastic model, ex post negotiations may no longer be optimal since they may induce a for-profit hospital to increase its debt-to-equity ratio, which in turn increases the hospital's risk of bankruptcy. Second, we have only considered a static model. In a dynamic model, insurers may try to behave opportunistically by renegotiating reimbursement rates. Martimort's (1994) model of a dynamic common agency suggests that a multi-payer system may be optimal if the hospital has private information concerning its costs.

APPENDIX

Proof of Proposition 1:

(1) The Single-Payer System. First note that the equilbrium T_S is the boundary solution $T_S = 0$. So the Lagrangian of the public payer's program is

$$\mathcal{L} = U - \alpha X - pT + \mu Y_I + \lambda Y.$$

The $Y \geq 0$ constraint will be nonbinding. We will later verify that indeed $\lambda = 0$ in equilibrium. Thus, the first order condition $\frac{\partial \mathcal{L}}{\partial \alpha} = 0$ provides $\mu = \frac{X}{X_I}$. Next,

$$\frac{\partial \mathcal{L}}{\partial I} = U_I - \alpha X_I + \frac{X}{X_I} Y_{II} + (1 - \frac{X X_{II}}{X_I^2}) = U_I - cX_I - c_I X + \frac{X}{X_I} Y_{II}. \tag{A1}$$

Thus, $I_S < I^E$. Next we verify that $\lambda = 0$. From $Y_I = 0$ we have $\alpha = \frac{cX_I + c_I X}{X_I}$. Therefore, $Y = \frac{c_I X^2}{X_I} > 0$. Hence, $\lambda = 0$.
(2) The Multi-Payer System. First, note that there exists a maximum $n_1 \in (0,1)$ such that for all $n \leq n_1$, the program $\max_T \Delta$ has an interior solution solving

$$\Delta_T = (1-n)U_T - c_T X + p - 1 = 0. \tag{A2}$$

For $n \geq n_1$, the equilibrium T_0 is the boundary solution $T_0 = 0$. For $n \leq n_1$, the Lagrangian for the public payer's program is

$$\mathcal{L} = n(U - \alpha X) - pT + \mu \Delta_I + \gamma \Delta_T + \lambda \Delta.$$

We will for now assume that $\lambda = 0$ and later verify this. The first order condition $\frac{\partial \mathcal{L}}{\partial \alpha} = 0$ provides $\mu = \frac{X}{X_I}$. Thus, $\frac{\partial \mathcal{L}}{\partial T} = 0$ reveals that

$$\Delta_{TT} \gamma = p - nU_T - \frac{X}{X_I} \Delta_{IT}. \tag{A3}$$

Next, from the first order condition $\frac{\partial \mathcal{L}}{\partial p} = 0$ we obtain $\gamma = T$ when $p_0 \in (0,1)$. In this case, equation (A3) reveals that

$$p_0 = T\Delta_{TT} + nU_T - \frac{X}{X_I} \Delta_{IT}. \tag{A4}$$

Define $K(I,T) = T\Delta_{TT} + \frac{X}{X_I} \Delta_{IT}$. Using equation (A4), equation (A2) becomes

$$\Delta_T(I,T) = U_T(I,T) - c_T X(I) - nU_T(I,T) + nU_T(I_0,T_0) + K(I_0,T_0). \tag{A5}$$

Next, note that at $(I_0, T^E(I_0))$, equation (A5) reduces to

$$\Delta_T(I_0, T^E(I_0)) = nU_T(I_0,T_0) - nU_T(I_0,T^E(I_0)) + K(I_0,T_0) \tag{A6}$$

when $p_0 \in (0,1)$. Next, we claim that Assumption 1 ($\epsilon_{TT} < \epsilon_{IT}$) implies that $K(I_0, T_0) > 0$. First, since $\epsilon_{TT} < \epsilon_{IT}$, we have $1 < \frac{\epsilon_{IT}}{\epsilon_{TT}} = \frac{-I\Delta_{IT}}{T\Delta_{TT}}$. Thus,

$$T\Delta_{TT} + I\Delta_{IT} > 0. \tag{A7}$$

However, by concavity of $X(I)$, we have $\frac{X}{X_I} > I$. Thus, equation (A7) implies that $K(I,T) = T\Delta_{TT} + \frac{X}{X_I}\Delta_{IT} > 0$. Next, we claim that $\Delta_T(I_0, T^E(I_0)) > 0$ (in equation (A6)) when $K(I_0, T_0) > 0$. Suppose instead that $\Delta_T(I_0, T^E(I_0)) < 0$ when $K(I_0, T_0) > 0$. Then, by concavity, $\Delta_T(I_0, T^E(I_0)) < 0$ implies that $T^E(I_0) > T_0(I_0)$. but, then $nU_T(I_0, T_0) - nU_T(I_0, T^E(I_0)) > 0$, implying that $K(I_0, T_0) < 0$ from equation (A6). This is a contradiction. Therefore, we must have $\Delta_T(I_0, T^E(I_0)) > 0$ when $K(I_0, T_0) > 0$ and $p_0 \in (0,1)$. That is, the hospital overinvests in technology when $K > 0$ and $p_0 \in (0,1)$. If $p_0 = 1$ for all n, then by the intermediate value theorem, there exists an $\hat{n} \in (0,1)$ such that $nU_T(I_0, T^E(I_0)) < 1$ for all $n < \hat{n}$. Thus,

$$\Delta_T(I_0, T^E(I_0)) = 1 - nU_T(I_0, T^E(I_0)) > 0$$

for all $n < \hat{n}$ when $p_0 = 1$. That is, the hospital overinvests in technology for $n < \hat{n}$ when $p_0 = 1$ for all n.
Now we define n_0 as the largest $n \in (0,1)$ such that

$$(I_0(n), T_0(n)) = (I_0(n), T^E(I_0(n))) \tag{A8}$$

From the above analysis, we know that n_0 exists by continuity. To show that overinvestment in technology occurs for all $n \leq n_0$, note that equation (A8) can hold only when $p_0 = 1$. For, using equation (A8), equation (A5) can be reduced to $\Delta_T(I_0(n_0), T^E(I_0(n_0))) = K > 0$ when $p_0 \in (0,1)$. However, this is a contradiction since $\Delta_T = 0$ at T^E. Thus, $p_0 = 1$ at n_0. Moreover, $n_0 U_T(I_0, T^E(I_0)) = 1$, implying that $\hat{n} = n_0$. Therefore, for $n < n_0$, either $p_0 = 1$ and $T_0 > T^E$ or $p_0 \in (0,1)$ with $T_0 \geq T^E$.
Next, we claim that $I_0 < I^E$ when $K(I_0, T_0) > 0$. For $n \geq n_1$, $I_0 < I^E$ from equation (A1). When $n \leq n_1$ and $p_0 \in (0,1)$,

$$\frac{\partial \mathcal{L}}{\partial I} = n(U_I - \alpha X_I) + \mu_I \Delta_I + \mu \Delta_{II} + \gamma \Delta_{IT} = (U_I - c_I X - c X_I) + \frac{X}{X_I}\Delta_{II} + T\Delta_{IT} = 0 \tag{A9}$$

since $\gamma = T$ from the first order condition for p. Define $L(I,T) = \frac{X}{X_I}\Delta_{II} + T\Delta_{IT}$. Next, note that $K(I,T) > 0$ implies that

$$\frac{TX_I}{X}\Delta_{TT} > -\Delta_{IT} \tag{A10}$$

and that $L(I,T) > 0$ implies that

$$\frac{X}{TX_I}\Delta_{II} > -\Delta_{IT}. \tag{A11}$$

Thus, multiplying equations (A10) and (A11) together provides $\Delta_{TT}\Delta_{II} < \Delta_{IT}^2$. However, this violates the hospital's second order condition $\Delta_{TT}\Delta_{II} - \Delta_{IT}^2 > 0$. Therefore, to satisfy this second order condition, we must have $L(I,T) < 0$ when $K(I,T) > 0$ and $p_0 \in (0,1)$. As a result, equation (A9) reveals that $\frac{\partial \mathcal{L}}{\partial I} = L(I^E(T_0), T_0) < 0$ when $p_0 \in (0,1)$. Thus, the hospital undersupplies intensity. A similar proof holds when $p_0 = 1$.

Finally, we verify that $\lambda = 0$ under Assumption 2 ($\frac{U_I}{U} < \frac{X_I}{X}$). We must show that $\Delta(I_0, T_0) \geq 0$ for all n. Define $\hat{\alpha}$ to be the maximum out-of-equilibrium α that forces $\Delta = 0$ for a fixed (I,T) when $p = 1$. Then $n\hat{\alpha} = c - (1-n)\frac{U}{X}$, while the equilibrium $n\alpha_0 = c_I\frac{X}{X_I} + c - (1-n)\frac{U_I}{X_I}$. Thus, $\alpha_0 \geq \hat{\alpha}$ at (I_0, T_0) if $c_I\frac{X}{X_I} > (1-n)[\frac{U_I}{X_I} - \frac{U}{X}]$, which holds if $\frac{U_I}{U} \leq \frac{X_I}{X}$. Thus, $\Delta(I_0, T_0) \geq 0$ and $\lambda = 0$ under Assumption 2.

(3) The All-Payer System. The first order condition for the Nash bargaining program $\max_\phi Y^q V^x M^z$ is

$$qMV = nzVY + (1-n)xMY. \tag{A12}$$

Fron equation (A12) we can derive $\phi X = qU + (x+z)(cX+T)$. Therefore,

$$Y = \phi X - cX - T = qU + (x+z-1)(cX+T) = q(U - cX - T)$$

since $q + x + z = 1$. Therefore, $\frac{\partial Y}{\partial T} = 0$ and $\frac{\partial Y}{\partial I} = 0$ at (I^E, T^E).

Proof of Proposition 2:

(1) Single-Payer System. Note that the public reimbursement plan ($\alpha = 0, r = 1, p = 1$) would ensure that $Y = 0$ for any I and T. Since the hospital is indifferent, (I^E, T^E) is an equilibrium.

(2) Multi-Payer System. First note that if $r > 0$, then

$$\Delta_T = U_T - c_T X - 1 + p - nU_T + nrc_T X. \tag{A13}$$

From equation (A2), we see that equation (A13) reduces to $\Delta_T = nrc_T X > 0$ at $T_0(I)$. Thus, $T_1 > T_0$. To verify that $r > 0$, we examine the public payer's Lagrangian:

$$\mathcal{L} = n(U - \alpha X) - nrcX - pT + \mu\Delta_I + \gamma\Delta_T + \lambda\Delta.$$

Assume for now that $\lambda = 0$. The first order condition for α implies that $\mu = \frac{X}{X_I}$. Thus, the first order condition for p then provides $\gamma = T$ if $p \in (0,1)$. Next,

$$\frac{\partial \mathcal{L}}{\partial r} = -ncX + \frac{nX}{X_I}[c_I X + cX_I] + nc_T XT > 0$$

implies that $r = 1$ in equilibrium. Next, the first order condition on I is equation (A9) above. Thus, $I_1(T) = I_0(T)$ for all T when $p_0, p_1 \in (0,1)$.

Proof of Proposition 3:

First, note that the total surplus to be bargained over now becomes

$$\Delta = (1-n)U - cX + n\alpha X + \psi(G) + (p-1)(T+G).$$

The government's program is now

$$\max_{\alpha,p,G,T} n(U - \alpha X) - p(T + G) + \mu\Delta_I + \gamma\Delta_T + \lambda\Delta_G.$$

Differentiating with respect to α provides $\mu = \frac{X}{qX'}$. Differentiation with respect to G reveals that $\lambda = \frac{p}{\psi''(G)}$. The first order condition for T is still equation (A3). Next, solving for γ in equation (A3) and substituting into the first order condition for p

$$-(T + G) + \gamma + \frac{p}{\psi''(G)}$$

when $p \in (0, 1)$ provides the equilibrium pass-through payment

$$p = \frac{\Delta_{TT}\psi''}{\Delta_{TT} + \psi''}\left[T + G + \frac{(nU_T + \frac{X}{X_I}\Delta_{TI})}{\Delta_{TT}}\right]. \tag{A14}$$

Using the the definition of p_0 in equation (A4), equation (A14) reduces to

$$p = \frac{\psi''}{\Delta_{TT} + \psi''}[p_0 + G\Delta_{TT}]. \tag{A15}$$

However, since $\Delta_G = 0$ in equilibrium, we must also have $p = 1 - \psi'(G)$. Thus, in equilibrium, equation (A15) implies that

$$1 - \psi'(G) = \frac{\psi''(G)}{\Delta_{TT} + \psi''(G)}[p_0 + G\Delta_{TT}]. \tag{A16}$$

Letting $\psi(G) = aG - \frac{b}{2}G^2$, it is straightforward to show that $G^* = \frac{p_0}{b - 2\Delta_{TT}} > 0$ solves equation (A16) and reduces equation (A15) to $p = bG^*$. □

Proof of Proposition 5:

Suppose that $D^* = 0$. Then z^*, $z^{**} = 0$. First, we claim that $\frac{\partial\beta^*}{\partial D} > 0$. Let $H = Y^qV^{1-q}$. The private insurer will choose β to maximize H by solving the first order condition $qV(1 + b\frac{\dot{X}}{X}) = (1 - q)Y$. Next note that

$$\frac{\partial\beta^*}{\partial D} = \frac{-H_{\beta D}}{H_{\beta\beta}}. \tag{A17}$$

By concavity of H, $H_{\beta\beta} < 0$. Moreover,

$$H_{\beta D} = qV\frac{\partial^2 Y}{\partial\beta\partial D} + (1 - q)\frac{\partial Y}{\partial D}\frac{\partial V}{\partial\beta}.$$

However, since $z^* = z^{**} = 0$, we have $\frac{\partial Y}{\partial D} = -bF(z^*) = 0$ and

$$\frac{\partial^2 Y}{\partial\beta\partial D} = -b\frac{\partial F(z^*)}{\partial z}\frac{\partial z^*}{\partial\beta} > 0.$$

Thus, $H_{\beta D} > 0$. Hence, $\frac{\partial \beta^*}{\partial D} > 0$. Now now that

$$\frac{dY}{dD} = (1-n)X\frac{\partial \beta}{\partial D} - bF(z^*) + b(1-n)\int_0^{z^*} X\frac{\partial \beta}{\partial D}dF. \tag{A18}$$

Since we assume $D^* = 0$, the hospital's first order condition must be $\frac{dY}{dD} \leq 0$. However, since $z^* = 0$, we must also have

$$\frac{dY}{dD} = (1-n)X\frac{\partial \beta^*}{\partial D} > 0$$

from equation (A18). However, this is a contradiction. Therefore, we must have $D^* > 0$.□

Proof of Theorem 1:

We will derive the equilibrium form (4-2) in the following manner. First, we conjecture that A_2 takes this form in equilibrium. Then we show that M_1's best response also takes the quadratic form in (4-2). To do so, the following Lemma is useful.

Lemma A *Given the conjecture A_2 as defined in (4-2), the agent's equilibrium utility in a market of size θ is $U(\theta) = w_2(\theta) + \max_{x_1,p_1}(\mathcal{K}(x_1,p_1,\theta) - A_1(x_1,p_1))$, where*

$$w_2(\theta) = \max_{x_2,p_2}\Big\{(1-h_2-\sigma_2)x_2p_2 - \frac{a_2+\beta_2}{2}x_2^2 -$$

$$-\frac{b_2+\eta_2}{2}p_2^2 + (\theta-\alpha_2)x_2 + (\theta-\delta_2)p_2 - \gamma_2\Big\} \text{ and}$$

$$\mathcal{K}(x_1,p_1,\theta) = \frac{k_1}{2}x_1^2 + \frac{k_2}{2}p_1^2 + (k_3+\theta\phi_1)x_1 + (k_4+\theta\phi_2)p_1 + k_5x_1p_1 \text{ for}$$

$$\phi_i = 1 + ct_i + cs_it_i \text{ ,i=1,2,}$$

$$k_1 = -a_1 + c^2s_1t_1$$

$$k_2 = -b_1 + c^2t_2$$

$$k_3 = -ct_1(\delta_2 + \alpha_2 s_1)$$

$$k_4 = -ct_2(\delta_2 + \alpha_2 s_2)$$

$$k_5 = (1-h_1) + c^2t_1$$

$$t_1 = \frac{(1-h_2-\sigma_2)}{(a_2+\beta_2)(b_2+\eta_2) - (1-h_2-\sigma_2)^2}$$

$$t_2 = \frac{(a_2+\beta_2)}{(a_2+\beta_2)(b_2+\eta_2) - (1-h_2-\sigma_2)^2}$$

$$s_1 = \frac{b_2+\eta_2}{1-h_2-\sigma_2} \text{ and } s_2 = \frac{1-h_2-\sigma_2}{a_2+\beta_2}.$$

The proof is tedious, and so is omitted. The details are available upon request.

From the Revelation Principle, for any tax $A_1 : \Re^2_+ \to \Re_+$ that is a best response to the conjecture A_2, there exists a direct revelation vertical restraint $(p_1(\theta), x_1(\theta), T_1(\theta))$ such that $T_1(\theta) = A_1(x_1(\theta), p_1(\theta))$ when restricted to the domain $\mathcal{D} \equiv \{(x_1(\theta), p_1(\theta)) \mid \theta \in \Theta\}$ and $(x_1(\theta), p_1(\theta), T_1(\theta))$ solves the program

$$[\tilde{M_1}] \begin{cases} \max_{p_1, x_1, T_1} \int_\Theta \{T_1(z) - m_1 x_1(z)\} f(z) dz \text{ such that} \\ U(\theta,\theta) \geq U(\hat{\theta},\theta), \\ U(\theta,\theta) \geq \max\{0, w_2(\theta)\}, \text{ and} \\ p_1, x_1 \geq 0, \text{ for all } \theta, \hat{\theta} \in \Theta, \text{ } where \end{cases}$$

$$U(\hat{\theta},\theta) \equiv w_2(\theta) + \frac{k_1}{2} x_1(\hat{\theta})^2 + \frac{k_2}{2} p_1(\hat{\theta})^2 + (k_3 + \theta\phi_1) x_1(\hat{\theta}) +$$
$$+ (k_4 + \theta\phi_2) p_1(\hat{\theta}) + k_5 x_1(\hat{\theta}) p_1(\hat{\theta}) - T_1(\hat{\theta}).$$

It is important to note that the Revelation Principle does not in general hold simultaneously for both manufacturers unless we restrict reports of θ to Θ (see Martimort and Stole [1993]). Next, let $u(\theta) \equiv U(\theta,\theta) - w_2(\theta)$ be the rent function corresponding to A_1 and suppose that that (x_1, p_1) is a solution to the program

$$[M_1] \begin{cases} \max_{p_1, x_1, u} \int_\Theta \{T_1(z) - m_1 x_1(z)\} f(z) dz \text{ such that} \\ u'(\theta) = \phi_1 x_1(\theta) + \phi_2 p_1(\theta), \\ u(\underline{\theta}) \geq \max\{0, -w_2(\underline{\theta})\}, \text{ and} \\ p_1, x_1 \geq 0 \text{ for all } \theta, \hat{\theta} \in \Theta. \end{cases}$$

Then (x_1, p_1) will also be an optimum of the program $[\tilde{M_1}]$ iff $\phi_1 x_1(\theta) + \phi_2 p_1(\theta)$ is strictly increasing in θ. This conclusion follows from Theorems 7.1 and 7.3 of Fudenberg and Tirole [1991]. Since $-w_2(\theta)$ is decreasing and $u(\theta)$ is increasing, the participation constraint of program $[M_1]$ holds under assumptions (A22) and (A23) below. In fact, if $\phi_i \geq 0$, i=1,2, then program $[\tilde{M_1}]$ is equivalent to program $[M_1]$ if x_1 and p_1 are strictly increasing. We will assume *a priori* that ϕ_1 and ϕ_2 are nonnegative and then check *ex post* to see when this is indeed the case at the identified solution. Also, we will assume the following:

$(A19)$ $a_2 + \beta_2 > 0$, $b_2 + \eta_2 > 0$, and $(a_2 + \beta_2)(b_2 + \eta_2) - (1 - h_2 - \sigma_2)^2 > 0$;

$(A20)$ $\underline{\theta}(\phi_i - 1) \geq \frac{k_{2+i}}{c}$, $i = 1, 2$;

$(A21)$ $k_1, k_2 < 0$, and $k_1 k_2 - k_5^2 > 0$; and

$(A22)$ $\underline{\theta} - \frac{1}{f(\underline{\theta})} \geq \max\left\{ \frac{k_3 - \frac{k_1 k_4}{k_5} - m_1}{\phi_1 - \frac{k_1 \phi_2}{k_5}}, \frac{k_3 - \frac{k_5 k_4}{k_2} - m_1}{\phi_1 - \frac{k_5 \phi_2}{k_2}} \right\}.$

$(A23)$ $x_1(\theta)p_1(\theta) - e(x_1(\theta), x_2(\theta), p_1(\theta), p_2(\theta), \theta) - m_1x_1 \geq -w_2(\theta)$ at $\theta = \underline{\theta}$, where $x_1(\theta)$ and $p_1(\theta)$ are given below and $x_2(\theta)$ and $p_2(\theta)$ optimize $w_2(\theta)$.

Conditions (A19) and (A20) insure that $w_2(\theta)$ is the optimum of a concave program, attained at a positive level of sales. Similarly, $[M_1]$ is a concave program if the inequalities in (A21) hold. Under (A22) and (A23), the agent will be employed by manufacturer M_1 for any realization of the demand. Clearly, all five assumptions must be checked *ex post* since they involve the equilibrium parameters of A_2. The Hamiltonian for program $[M_1]$ is

$$H(u,p,x,\theta) = f(\theta)[\mathcal{K}(x_1,p_1,\theta) - u(\theta) - m_1x_1] + \mu(\theta)[\phi_1x_1 + \phi_2p_1], \tag{A24}$$

where u is the state variable. The Pontryagin principle yields $\dot{\mu}(\theta) = f(\theta)$. Furthermore, $\overline{\theta}$ is a free boundary so that $\mu(\overline{\theta}) = 0$. Integrating provides $\mu(\theta) = F(\theta) - 1$. Substituting this back into the Hamiltonian, we have

$$\frac{1}{f(\theta)}\frac{\partial H}{\partial x_1} = k_1x_1 + k_3 + \theta\phi_1 + k_5p_1 - m_1 - \frac{1-F(\theta)}{f(\theta)}\phi_1 = 0 \text{ and} \tag{A25}$$

$$\frac{1}{f(\theta)}\frac{\partial H}{\partial p_1} = k_2p_1 + k_4 + \theta\phi_2 + k_5x_1 - \frac{1-F(\theta)}{f(\theta)}\phi_2 = 0. \tag{A26}$$

Solving (A25) and (A26) provides

$$x_1(\theta) = \left(\frac{k_2}{k_5^2 - k_2k_1}\right)\left[\left(\theta - \frac{1-F(\theta)}{f(\theta)}\right)\left(\phi_1 - \frac{k_5\phi_2}{k_2}\right) + k_3 - \frac{k_5k_4}{k_2} - m_1\right], \tag{A27}$$

$$p_1(\theta) = \left(\frac{-k_5}{k_5^2 - k_2k_1}\right)\left[\left(\theta - \frac{1-F(\theta)}{f(\theta)}\right)\left(\phi_1 - \frac{k_1\phi_2}{k_5}\right) + k_3 - \frac{k_1k_4}{k_5} - m_1\right], \tag{A28}$$

and $T_1(\theta) = \mathcal{K}(\theta) - \int_{\underline{\theta}}^{\theta} \{\phi_1x_1(z) + \phi_2p_1(z)\}dz - \max\{0\ ,\ -w_2(\underline{\theta})\}$, where

$$w_2(\underline{\theta}) = -\gamma_2 - \frac{a_2+\beta_2}{2}x_0^2 - \frac{b_2+\eta_2}{2}p_0^2 + (\underline{\theta} - \alpha_2)x_0 + (\underline{\theta} - \delta_2)p_0 + (1 - h_2 - \sigma_2)x_0p_0$$

$$for \;\; x_0 \equiv \frac{\underline{\theta}(\phi_1 - 1) + k_3}{c} \;\; and \;\; p_0 \equiv \frac{\underline{\theta}(\phi_2 - 1) + k_4}{c}.$$

It is immediate from (A27) and (A28) that x_1 and p_1 will be increasing in θ if the inverse hazard rate of $f(\theta)$ is weakly decreasing when $k_2\phi_1 - k_5\phi_2 \leq 0$ and $k_5\phi_1 - k_1\phi_2 \geq 0$. Note that these last two inequalities hold when $k_5 \geq 0$. Finally, from Lemma A, it is evident that $k_5 \geq 0$ if $1 \geq h_2 + \sigma_2$, since we assumed that $1 \geq h_1$ (we verify below that $1 = h_2 + \sigma_2$ in equilibrium).

Next we show that the best response to A_2 is of form (4-2) for all $(x_1, p_1) \in \mathcal{D}$. For the Beta$(1, \frac{1}{\lambda})$ density, $\frac{1-F(\theta)}{f(\theta)} = \lambda(\overline{\theta} - \theta)$. Thus,

$$x_1(\theta) = \left(\frac{k_2}{k_5^2 - k_2k_1}\right)\left[\left((1+\lambda)\theta - \lambda\overline{\theta}\right)\left(\phi_1 - \frac{k_5\phi_2}{k_2}\right) + k_3 - \frac{k_5k_4}{k_2} - m_1\right]. \tag{A29}$$

Solving (A29) for θ provides

$$\theta = \left(\frac{k_5^2 - k_2 k_1}{k_2}\right) \frac{x_1}{(1+\lambda)(\phi_1 - \frac{k_5}{k_2}\phi_2)} + \frac{\lambda}{1+\lambda}\overline{\theta} -$$
$$-\frac{1}{1+\lambda}\left(\frac{1}{\phi_1 - \frac{k_5}{k_2}\phi_2}\right)\left(k_3 - \frac{k_5 k_4}{k_2} - m_1\right). \qquad \text{(A 30)}$$

In terms of p_1,

$$\theta = \left(\frac{k_5^2 - k_2 k_1}{-k_5}\right) \frac{p_1}{(1+\lambda)(\phi_1 - \frac{k_1}{k_5}\phi_2)} + \frac{\lambda}{1+\lambda}\overline{\theta} -$$
$$-\frac{1}{1+\lambda}\left(\frac{1}{\phi_1 - \frac{k_1}{k_5}\phi_2}\right)\left(k_3 - \frac{k_1 k_4}{k_5} - m_1\right). \qquad \text{(A 31)}$$

(A30) and (A31) can be rewritten as

$$\theta = \frac{1}{1+\lambda}[yAx_1 + \lambda\overline{\theta} - AC] \text{ and} \qquad \text{(A32)}$$

$$\theta = \frac{1}{1+\lambda}[zBp_1 + \lambda\overline{\theta} - BD]. \text{ Then}$$

$$T_1(\theta) = \left(\frac{k_1}{2} + \frac{\phi_1 yA}{2(1+\lambda)}\right)x_1^2 + \left(\frac{k_2}{2} + \frac{\phi_2 zB}{2(1+\lambda)}\right)p_1^2 +$$
$$+\left(k_3 + \frac{\phi_1(\lambda\overline{\theta} - AC)}{1+\lambda}\right)x_1 + \left(k_4 + \frac{\phi_2(\lambda\overline{\theta} - BD)}{1+\lambda}\right)p_1 + k_5 x_1 p_1 +$$
$$+\left[\frac{\phi_1(\lambda\overline{\theta} - AC)^2}{2yA(1+\lambda)} + \frac{\phi_2(\lambda\overline{\theta} - BD)^2}{2zB(1+\lambda)} - \frac{\phi_1\underline{\theta}(\lambda\overline{\theta} - AC)}{yA} - \frac{\phi_2\underline{\theta}(\lambda\overline{\theta} - BD)}{zB} +\right.$$
$$\left. + \frac{\phi_1(1+\lambda)\underline{\theta}^2}{2yA} + \frac{\phi_2(1+\lambda)\underline{\theta}^2}{2zB} - \max\{0, -w_2(\theta)\}\right]. \qquad \text{(A33)}$$

(A32) provides $\lambda\overline{\theta} - AC = (1+\lambda)\underline{\theta} - yA\underline{x}_1$. This allows a simplification of (A33) so that we the tax parameters can be rewritten as

$$\alpha_1 = k_3 + \phi_1\underline{\theta} - \underline{x}_1\Gamma_1, \quad \delta_1 = k_4 + \phi_2\underline{\theta} - \underline{p}_1\Gamma_2,$$

$$\beta_1 = k_1 + \Gamma_1, \quad \eta_1 = k_2 + \Gamma_2, \text{ and } \sigma_1 = k_5 \text{ where}$$

$$\Gamma_1 \equiv \frac{\phi_1}{1+\lambda}\left(\frac{k_5^2 - k_1 k_2}{k_2\phi_1 - k_5\phi_2}\right), \ \Gamma_2 \equiv \frac{\phi_2}{1+\lambda}\left(\frac{k_5^2 - k_1 k_2}{k_1\phi_2 - k_5\phi_1}\right),$$

$$\underline{x}_1 = \frac{\phi_1}{\Gamma_1}\left(\underline{\theta} + \frac{\lambda\overline{\theta}}{1+\lambda}\right) + \frac{k_2 k_3 - k_4 k_5 - k_2 m_1}{k_5^2 - k_1 k_2},$$

$$and\ \underline{p}_1 = \frac{\phi_2}{\Gamma_2}\left(\underline{\theta} + \frac{\lambda\overline{\theta}}{1+\lambda}\right) - \frac{k_5k_3 - k_4k_1 - k_5m_1}{k_5^2 - k_1k_2}.$$

The conclusion that $\sigma_1 = k_5$ warrants special attention. Recall that by definition $k_5 = 1 - h_1 + c^2t_1$. Now note that from Lemma A, if we conjecture that $\sigma_2 = 1 - h_2$, then $t_1 = 0$, resulting in $\sigma_1 = k_5 = 1 - h_1$. Thus, $\sigma_i = 1 - h_i$, i=1,2, is a royalty rate that can indeed be sustained in equilibrium. Finally, note that $\phi_i \geq 0$ in equilibrium, i=1,2, if $c \geq -\max\{(b_i+\eta_i)\ ,\ (a_i+\beta_i) \mid i = 1,2\}$ Note that this maximum is negative by assumption (A19). Thus $\phi_i \geq 0$ at least for complements and weak substitutes.

Next, we solve for the fixed franchise fees. From $T_1(\theta)$ above, it can be show that the fixed franchise fees are

$$\gamma_1 = \frac{\Gamma_1\underline{x}_1^2}{2} + \frac{\Gamma_2\underline{p}_1^2}{2} \text{ and } \gamma_2 = \frac{\tilde{\Gamma}_1\underline{x}_2^2}{2} + \frac{\tilde{\Gamma}_2\underline{p}_2^2}{2},$$

if $w_2(\underline{\theta}) \geq 0$. We will show that $w_2(\underline{\theta}) \geq 0$ iff $c \leq 0$.

First, note that γ_2 is chosen as large as possible by M_2 so that as much rent is extracted while still ensuring that the agent's rent in common agency is not below the level of rent the agent would obtain for exclusively serving M_1. At $\underline{\theta}$ this constraint is binding, i.e.,

$$\max_{x_1x_2p_1p_2}\{x_1p_1 + x_2p_2 - e(x_1,p_1,x_2,p_2,\underline{\theta}) - A_1 - A_2\} =$$

$$= \max_{x_1p_1}\{x_1p_1 - e(x_1,p_1,0,0,\underline{\theta}) - A_1\}. \tag{A34}$$

(A34) isolates

$$\gamma_2 = \max_{x_1x_2p_1p_2}\{x_1p_1 + x_2p_2 - e(x_1,p_1,x_2,p_2,\underline{\theta}) - \alpha_1x_1 - \delta_1p_1 -$$

$$-\frac{\beta_1}{2}x_1^2 - \frac{\eta_1}{2}p_1^2 - \sigma_1x_1p_1 - \alpha_1x_2 - \delta_2p_2 - \frac{\beta_2}{2}x_2^2 - \frac{\eta_2}{2}p_2^2 - \sigma_2x_2p_2\} -$$

$$-\max_{x_1p_1}\{x_1p_1 - e(x_1,p_1,0,0,\underline{\theta}) - \alpha_1x_1 - \delta_1p_1 - \frac{\beta_1}{2}x_1^2 - \frac{\eta_1}{2}p_1^2 - \sigma_1x_1p_1\}. \tag{A35}$$

Refering to Lemma A, we can define $\mathcal{K}_1$ and $\mathcal{K}_2$ so that $w_2(\underline{\theta}) = -\gamma_2 + \mathcal{K}_2$ and so that $\mathcal{K}_1$ is the value of the second maximization in the (A35). Then $w_2(\underline{\theta}) \geq 0$ iff $\mathcal{K}_1 + \mathcal{K}_2 \geq \max_{x_1p_1x_2p_2}\{L\}$, where L is the first maximand in equation (4-3). Next, we have $L \leq L - cx_1x_2 - cp_1p_2$ iff $c \leq 0$. However, note that $\mathcal{K}_1 + \mathcal{K}_2 = \max_{x_1x_2p_1p_2}\{L - cx_1x_2 - cp_1p_2\}$. Therefore, we indeed have $\mathcal{K}_1 + \mathcal{K}_2 \geq \max_{x_1x_2p_1p_2}\{L\}$. Hence, $w_2(\underline{\theta}) \geq 0$ iff $c \leq 0$.

Next, for $c > 0$, suppose that both participation constraints are binding:

$$\gamma_1 = \max_{x_1x_2p_1p_2}\{L\} - \mathcal{K}_2 \text{ and} \tag{A36}$$

$$\gamma_2 = \max_{x_1x_2p_1p_2}\{L\} - \mathcal{K}_1. \tag{A37}$$

Then $-\gamma_1 - \gamma_2 + \max_{x_1x_2p_1p_2}\{L\} = \mathcal{K}_1 + \mathcal{K}_2 - \max_{x_1x_2p_1p_2}\{L\}$. However, the right-hand side is negative if $c > 0$. Since the left-hand side is the rent the agent receives under common agency, the individual rationality constraint is violated if both participation constraints (A36) and (A37) bind. Consequently, under complements we must have $-\gamma_1 - \gamma_2 + \max_{x_1x_2p_1p_2}\{L\} = 0$ with $\gamma_i \leq \max_{x_1x_2p_1p_2}\{L\} - \mathcal{K}_{-i}$, i=1,2.

Finally, we must show that A_1 can be extended from $\mathcal{D}$ to $\Re_+^2$ with the same quadratic specification (4-2). For any x_1 $p_1 \in \Re_+$, let $I(x_1,p_1|\theta)$ be the tax level for which the retailer's utility from $(I(x_1,p_1|\theta),x_1,p_1)$ is $u(\theta)$. That is, the retailer is indifferent between choosing either $(A_1,x_1(\theta),p_1(\theta))$ and $(I(x_1,p_1|\theta),x_1,p_1)$, for any x_1 and p_1. To show that A_1 is implementable on $\Re_+^2$, it suffices to show that $I(x_1,p_1|\theta) < A_1(x_1,p_1)$ for all x_1, $p_1 \in \Re_+$ and all θ, with equality only at $x_1 = x_1(\theta)$ and $p_1 = p_1(\theta)$, where $x_1(\theta)$ and $p_1(\theta)$ are equilibrium choices. From Lemma A, we can write $I(x_1,p_1|\theta) = \mathcal{K}(\theta) - u(\theta)$. Then

$$A_1(x_1,p_1) - I(x_1,p_1|\theta) = \phi_1(\underline{\theta}-\theta)(x_1 - x_1(\theta)) + \underline{x}_1\Gamma_1(x_1(\theta)-x_1)+$$

$$+\frac{\Gamma_1}{2}(x_1^2 - x_1^2(\theta)) + \phi_2(\underline{\theta}_\theta)(p_1 - p_1(\theta)) + \underline{p}_1\Gamma_1(p_1(\theta)-p_1) + \frac{\Gamma_2}{2}(p_1^2 - p_1^2(\theta)).$$

Note that indeed $A_1(x_1(\theta),p_1(\theta)) - I(x_1,p_1|\theta) = 0$. Next,

$$\frac{\partial(A_1(x_1,p_1) - I(x_1,p_1|\theta))}{\partial x_1} = \phi_1(\underline{\theta}-\theta) + \Gamma_1(x_1 - \underline{x}_1) = 0$$

for $x_1^{**} \equiv \frac{\phi_1}{\Gamma_1}(\underline{\theta}-\theta) + \underline{x}_1$. However, we claim that $x_1^{**} = x_1(\theta)$. Note that

$$x_1^{**} - x_1(\theta) = \frac{\phi_1}{\Gamma_1}(\theta - \underline{\theta}) + \underline{x}_1 - \frac{\phi_1}{\Gamma_1}\left(\theta + \frac{\lambda\overline{\theta}}{1+\lambda}\right) - \frac{k_2k_3 - k_4k_5 - k_2m_1}{k_5^2 - k_1k_2}$$

$$= \frac{-\phi_1}{\Gamma_1}(\underline{\theta} + \frac{\lambda\overline{\theta}}{1+\lambda}) - \frac{k_2k_3 - k_4k_5 - k_2m_1}{k_5^2 - k_1k_2} + \underline{x}_1 = -\underline{x}_1 + \underline{x}_1 = 0.$$

Similarly,

$$\frac{\partial(A_1(x_1,p_1) - I(x_1,p_1|\theta))}{\partial p_1} = 0$$

at $p_1(\theta)$. Note also that $A_1(x_1,p_1) - I(x_1,p_1|\theta)$ is convex. Thus, $A_1(x_1,p_1) - I(x_1,p_1|\theta) > 0$, for all $(x_1,p_1) \neq (x_1(\theta),p_1(\theta))$. □

Proof of Proposition 6 and Corollary 3:

Let H^D and H^M be the Hamiltonians for the duopoly and monopoly cases. Then

$$\frac{\partial H^D}{\partial x_1} = \frac{\partial H^M}{\partial x_1} + \frac{c}{a_2+\beta_2}\left[-(a_2+\beta)x_2 + cx_1 - \alpha_1 + \left(\theta - \frac{1-F(\theta)}{f(\theta)}\right)\right]$$

$$= \frac{\partial H^M}{\partial x_1} + \frac{c}{a_2+\beta_2}\left[\frac{\partial U^D}{\partial x_2} - \frac{1-F(\theta)}{f(\theta)}\right],$$

from Lemma A. However, $\frac{\partial U^D}{\partial x_2} = \frac{\partial H^D}{\partial x_1} = 0$ in equilibrium. Thus, $\frac{\partial H^M(x_1,p_1)}{\partial x_1} < (>)\ 0$ iff $c < (>)\ 0$, since $a_2 + \beta_2 > 0$ by assumption (A1). Therefore, $x_1^D > (<) x_1^M(p_1^D)$ iff $c < (>) 0$. Similarly, $p_1^D > (<) p_1^M(x_1^D)$ iff $c < (>) 0$. Next, note that $x_1^M(p_1)$ and $p_1^M(x_1)$ are both increasing (since $h_1 \leq 1$), regardless of c. Consequently, $x_1^D < (>) x_1^M$ and $p_1^D < (>) p_1^M$ iff $c > (<) 0$.

Next, let H^F be the Hamiltonian for the completely informed monopolist. Then $\frac{\partial H^M}{\partial x_1} = \frac{\partial H^F}{\partial x_1} - \frac{1-F(\theta)}{f(\theta)}$. Thus, $x_1^F(p_1^M) > x_1^M$, with equality at $\overline{\theta}$. Also,

$$\frac{\partial H^D}{\partial x_1} = \frac{\partial H^F}{\partial x_1} - \frac{1-F(\theta)}{f(\theta)}\left(1+\frac{c}{a_2+\beta_2}\right) = \frac{\partial H^F}{\partial x_1} - \frac{1-F(\theta)}{f(\theta)}\phi_1.$$

Hence, $x_1^D < x_1^F(p_1^D)$. Similarly, $p_1^D < p_1^F(x_1^D)$. Thus, duopoly involves quantity rationing and price ceilings. □

Proof of Proposition 7 and Corollary 4:

First, note that program [ED] is equivalent to the following program:

$$\begin{cases} \max_{p_1,x_1,\hat{\theta}}\Big(\int_{\underline{\theta}}^{\hat{\theta}}\{(p_1-m_1)x_1 - e(x_1,0,p_1,0,z) + F(z)(x_1+p_1-\Pi'_2(z))\}f(z)dz \\ + \int_{\hat{\theta}}^{\overline{\theta}}\{(p_1-m_1)x_1 - e(x_1,0,p_1,0,z) - (1-F(z))(x_1+p_1-\Pi'_2(z))\}f(z)dz\Big) \\ \text{such that} \\ (1)\ \ x_1(\theta)\le x_1(\hat{\theta})\ \ \textit{and}\ \ p_1(\theta)\le p_1(\hat{\theta})\ \ \textit{for all}\ \theta\in[\underline{\theta},\hat{\theta}], \\ (2)\ \ x_1(\theta)\ge x_1(\hat{\theta})\ \ \textit{and}\ \ p_1(\theta)\ge p_1(\hat{\theta})\ \ \textit{for all}\ \theta\in[\hat{\theta},\overline{\theta}]. \end{cases}$$

Necessary first order conditions are

$$G_1(x_1,p_1) \equiv p_1 - m_1 - a_1x_1 - h_1p_1 + \theta + \frac{F(\theta)}{f(\theta)} - \psi_1 = 0 \ \ \textit{and}$$

$$G_2(x_1,p_1) \equiv x_1 - b_1p_1 - h_1x_1 + \theta + \frac{F(\theta)}{f(\theta)} - \tau_1 = 0 \ \ \textit{for}\ \ \theta\in[\underline{\theta},\hat{\theta}]; \tag{A38}$$

$$p_1 - m_1 - a_1x_1 - h_1p_1 + \theta - \frac{1-F(\theta)}{f(\theta)} + \psi_2 = 0 \ \ \textit{and}$$

$$x_1 - b_1p_1 - h_1x_1 + \theta - \frac{1-F(\theta)}{f(\theta)} + \tau_2 = 0 \ \ \textit{for}\ \ \theta\in[\hat{\theta},\overline{\theta}],$$

where ψ_i and τ_i are the Lagrangian multipliers for constraint (i) above, i=1,2. Clearly, $\psi_i > 0$ and $\tau_i > 0$, i=1,2, on some common nondegenerate interval containing $\hat{\theta}$ (when $\hat{\theta}\in(\underline{\theta},\overline{\theta})$) or else one of the optimal controls will not be nondecreasing.

Finally, to prove Corollary 4, we show that the price floor increases if sales cannot be monitored. Without monitoring the agent will choose $x_1^*(p_1) = \frac{(1-h_1)}{a_1}p_1 + \frac{\theta-m_1}{a_1}$. Thus, $\frac{dx_1^*(p_1)}{dp_1} = \frac{1-h_1}{a_1} \ge 0$. So with the no-monitoring constraint binding, equation (A38) becomes

$$x_1^*(p_1) + (p_1-m_1)\left(\frac{1-h_1}{a_1}\right) - a_1x_1^*(p_1)\left(\frac{1-h_1}{a_1}\right) - b_1p_1 - h_1\left(\frac{1-h_1}{a_1}\right)p_1 -$$

$$-h_1 x_1^*(p_1) + \theta\left(\frac{1-h_1}{a_1}\right) + \theta + \frac{F}{f}\left(1 + \frac{1-h_1}{a_1}\right) = 0 \tag{A39}$$

on the non-pooling region. (A39) reduces to

$$G_2(x_1^*(p_1^N), p_1^N) + \left(\frac{1-h_1}{a_1}\right) G_1(x_1^*(p_1^N), p_1^N) = 0,$$

where p_1^N is optimal for this no-monitoring case. Hence, $G_2(x_1^*(p_1^N), p_1^N) < 0$, since $x_1^*(p_1^N) \leq x_1(p_1^N)$. Consequently, $p_1^N \geq p_1(x_1^*(p_1^N)) \geq p_1$, where the last inequality is due to $p_1(x_1)$ decreasing in x_1. □

REFERENCES

Aghion, P. and P. Bolton (1987), Contracts as a Barrier to Entry, *American Economic Review* 77, 388-401.

Banerji, S. and C. Simon (1992), Franchising vs. Ownership: A Contracting Explanation, mimeo, University of Chicago.

Bernheim, D. and Whinston, M.(1985), Common Marketing Agency as a Device for Facilitating Collusion, *RAND J. Econ.* 16, 269-281.
—(1986), Common Agency, *Econometrica* 54, 923-942.
—(1992), Exclusive Dealing, discussion paper no. 1622, Harvard Institute of Economic Research.

Besanko, D. and M. Perry (1993), Equilibrium Incentives for Exclusive Dealing in a Differentiated Products Oligopoly, *RAND Journal of Economics* 24, 646-667.

Biglaiser, G. and Mezzetti, C. (1993), Principals Competing for an Agent in the Presense of Adverse Selection and Moral Hazard, *J. Econ. Theory* 61, 302-330.

Blair, B. and Lewis,T. (1994), Optimal Retail Contracts with Asymmetric Information and Moral Hazard, mimeo, University of Florida, *RAND J. of Econ.*,forthcoming.

Clinton, William (1993), *The President's Health Security Plan*, Times Books.

Comanor, W. and H. Frech (1985), The Competitive Effects of Vertical Agreements?, *American Economic Review* 75, 539-546.

Diamond, Peter (1992), Organizing the Health Insurance Market, *Econometrica* 60, 1233-1254.

Dranove, David, Shanley, Mark, and Carol Simon (1992), Is Hospital Competition Wasteful? *RAND J. Econ.* 23, 247-62.

Dranove, David and William White (1994), Recent Theory and Evidence on Competition in Hospital Markets, *Journal of Economics and Management Strategy* 3, 169-210.

Eaton, J. and M. Engers (1990), Intertemporal Price Competition, *Econometrica* 58, 637-60.

Encinosa, William (1994), Optimal Hospital Capital Structure under Negotiated Prices, mimeo, University of Florida.

Encinosa, William and David Sappington (1993), Towards a Benchmark for Optimal Prudency Review, *Journal of Regulatory Economics*, forthcoming.

Evans, Robert (1986), Finding the Levers, Finding the Courage; Lessons from Cost Containment in North America, *Journal of Health Politics, Policy and Law* 11, 585-616.

Fraysse, J. (1993), Common Agency: Existence of an Equilibrium in the Case of Two Outcomes, *Econometrica* 61, 1225-29.

Fudenberg, D. and Tirole, J. (1991), *Game Theory*, MIT press.

Gal-Or, E. (1991a), A Commom Agency with Incomplete Information, *RAND J. Econ.* 22, 274-86.
—(1991b), Optimal Franchising in Oligopolistic Markets with Uncertain Demand, *International Journal of Industrial Organization* 9,365-388.
—(1994), Excessive Investment in Hospital Capacities, *Journal of Economics and Management Strategy* 3, 53-70.

Ginsburg, Paul (1988), Public Insurance Programs: Medicare and Medicaid, in *Health Care in America*, ed. Frech, H., 179-220.

Glazer, Jacob, and Thomas McGuire (1994), Payer Competition and Cost Shifting in Health Care, *Journal of Economics and Management Strategy* 3, 71-92.

Hackey, Robert (1993), New Wine in Old Bottles: Certificate-of-Need Enters the 1990s, *Journal of Health Politics, Policy and Law* 18, 927-35.

Hall,R. and C. Hitch (1939), Price Theory and Business Behavior, *Oxford Economic Papers* 2, 12-45.

Ippolito, P. (1988), *Resale Price Maintenance: Economic Evidence from Litigation.* Washington, DC, Federal Trade Commission.

Ivaldi, M. and Martimort, D. (1993), Competition under Nonlinear Pricing, mimeo, IDEI, Toulouse.

Katz, M. (1989), Vertical Contractual Relations, in R. Schmalensee and R. D. Willig, eds., *The Handbook of Industrial Organization*, New York: North Holland.

Kopit, W., and McCann, R. (1988), Towards a Definitive Anti-trust Standard for Non-profit Hospital Mergers, *Journal of Health Politics, Policy and Law* 13, 635-62.

Laffont, J. and Tirole, J. (1986), Using Cost Observation to Regulate Firms, *Journal of Political Economy* 94, 614-41.
—(1990), The Regulation of Multiproduct Firms, Part 2, *J. of Public Econ.* 43, 37-66.
—(1990a), Adverse Selection and Renegotiation in Procurement, *Review of Economic Studies* 57, 597-626.

Lafontaine, F. (1993), Contractual Arrangements as Signalling Devices: Evidence from Franchising, *Journal of Law, Economics, and Organization* 9(2), 256-89.

Leung, M. (1994), The Optimal Price Tax under Asymmetric Information, mimeo, The University of Western Ontario.

Lewis, T. and Sappington, D.(1988), Regulating a Monopolist with Unknown Demand, *American Economic Review* 78, 986-98.
—(1989a), Countervailing Incentives in Agency Problems, *J. Econ. Theory* 49, 294-313.
—(1989b), Inflexible Rules in Agency Problems, *American Economic Review* 79, 69-84.
—(1991), Technological Change and the Boundaries of the Firm, *American Economic Review*, 81, 887-900.
—(1992), Incentives for Conservation and Quality-Improvement by Public Utilities, *American Economic Review* 82, 1321-40.
—(1994), Insurance, Adverse selection, and Creamskimming, *Journal of Economic Theory*, forthcoming.

Ma, Ching-to Albert (1994), Health Care Payment Systems: Cost and Quality Incentives, *Journal of Economics and Management Strategy* 3, 93-112.

Ma, Ching-to Albert, and Thomas McGuire (1993), Paying for Joint Cost in Health Care, *Journal of Economics and Management Strategy* 2, 71-95.

Martimort, D. (1992), Multi-Principaux avec Anti-Selection, *Annales d'Economie et de Statistique* 28, 1-37.
—(1993), Exclusive Dealing, Common Agency, and Multiprincipal Incentive Theory, mimeo, IDEI, Toulouse.
—(1994),Multiprincipal Charters as a Safeguard Against Opportunism in Organizations, mimeo, IDEI and INRA-ESR.

Martimort, D. and L. Stole (1993), A Note on the Revelation Principle in Common Agency Games, mimeo.

Maskin, E. and J. Tirole (1988), A Theory of Dynamic Oligopoly, II: Price Competition, *Econometrica* 56, 571-600.

Mathewson, F. and R. Winter (1987), The Competitive Effects of Vertical Agreements: Comment, *American Economic Review* 77, 1057-62.

McManis, G. (1990), Competition's Failure Means It's Time for Collaboration, *Modern Health Care* 20, 57-58.

Merrill, Jeffrey (1994), *The Road to Health Care Reform*, Plenum Press.

Mezzetti, C. (1993), Common Agency with Horizontally Differentiated Principals: Regulating a Multinational Firm, mimeo, University of North Carolina.

Newhouse, Joseph (1988), Has Erosion of the Medical Marketplace Ended?, *Journal of Health Politics, Policy and Law* 3,, 263-78.
—(1994), Symposium on Health Care Reform, *Journal of Economic Perspectives* 8:3.

Perry, M. and Besanko, D. (1991), Resale Price Maintenance and Manufacturer Competition for Exclusive Dealerships, *Journal of Industrial Econ.* 39, 517-44.

Posner, R. (1977), The Rule of Reason and the Economic Approach: Reflections on the *Sylvania* Decision, *University of Chicago Law Review* 45, 1-20.

Prospective Payment Assessment Commission (1993), *Medicare and the American Health Care System: Report to Congress*, Washington, D.C.

Relman, Arnold (1989), Confronting the Crisis in Health Care, *Technology Review* 92, 30-40.

Rey, P. and J. Tirole. (1986), The Logic of Vertical Restraints, *American Economic Review* 76, 921-939.

Roberts, Marc (1993), *Your Money or Your Life: The Health Care Crisis Explained*, Doubleday.

Robinson, J. and Luft, H. (1985), The Impact of Hospital Market Structure on Patient Volume, Average Length of Stay, and the Cost of Care, *Journal of Health Economics* 4, 333-56.

Rogerson, William (1994), Choice of Treatment Intensity by a Nonprofit Hospital Under Prospective Pricing, *Journal of Economics and Management Strategy* 3, 7-51.

Romano, R. (1994), Double Moral Hazard and Resale Price Maintenance, *RAND Journal of Economics*, forthcoming.

Schwartz, M. (1987), The Competitive Effects of Vertical Agreements: Comment, *American Economic Review* 77, 1063-68.

Simpson, J. (1985), State Certificate-of-Need Programs: The Current Status, *American Journal of Public Health* 75, 1225-29.

Starr, Paul (1994), *The Logic of Health Care Reform*, Whittle Books.

Stigler, G. (1947), The Kinky Oligopoly Demand Curve and Rigid Prices, *Journal of Political Economy* 55, 442-444.

Stole, L. (1992), Mechanism Design under Common Agency, mimeo, University of Chicago.

Sweezy, P. (1939), Demand under Conditions of Oligopoly, *Journal of Political Economy* 47, 568-73.

Telser, L. (1960), Why Should Manufacturers Want Fair Trade?, *Journal of Law and Economics* 3, 86-105.

Thorpe, Kenneth (1993), The American States and Canada: A Comparative Analysis of Health Care Spending, *Journal of Health Politics, Policy and Law* 18, 477-489.

Tirole, J. (1988), *The Theory of Industrial Organization*, MIT Press.

BIOGRAPHICAL SKETCH

William Encinosa, III, was born in Tampa, Florida, in 1966. He received his B.A. in Mathematics at the University of South Florida in 1988, and his M.A. in Mathematics at the University of Florida in 1990. He entered the doctoral program in the Economics Department at the University of Florida in 1991, and expects to receive a Doctor of Philosophy degree in May 1995. Mr. Encinosa has accepted the position of Robert Wood Johnson Scholar in Health Policy at the University of Michigan-Ann Arbor.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

David E. M. Sappington, Chairman
Lanzillotti-McKethan Eminent Scholar
and Professor of Economics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Tracy Lewis
James Walters Eminent Scholar of
Entrepreneurship and Professor
of Economics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Jonathan Hamilton
Associate Professor of Economics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

David Brown
Nationsbank Associate Professor
of Finance, Insurance, and
Real Estate

This dissertation was submitted to the Graduate Faculty of the Department of Economics in the College of Business Administration and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

May 1995

Dean, Graduate School